明細書

類似率算出装置並びに類似率算出プログラム

技術分野

本発明は、技術文献群同士を比較して類似性を判断する類似率算出装置並びに類似率算出プログラムに関する。

背景技術

従来のパテントマップでは、特許文献を用いて、同一又は類似の研究開発テーマの内容について、技術比較を行って、全体的な動向、分布を知ることが可能であるとされている。そして経営者がパテントマップ見ることによって、市場動向、技術動向、参入企業及びライバル企業動向、将来性等の経営判断的要素を分析することが可能とされている。

またパテントマップでは、A社に関連した技術文献A群とB社に関連した技術文献B群とでマクロ的な比較を行う必要がある場合に、技術文献A群と技術文献B群に所属する個々の技術文献どうしをミクロ的に比較し、そこからマクロ的に技術文献群間の比較を導き出していた。

図１９は、技術文献A群に含まれる技術文献と、技術文献B群に含まれる技術文献を個々にミクロ的に比較する、従来の比較状況を示す図である。

図１９に示すように、技術文献A群に記載されている技術に対し、比較対象の技術文献B群の技術とを比較する場合、従来は、技術文献A群に含まれる技術文献（特許公報や技報など）と技術文献B群に含まれる技術文献（特許公報や技報など）を総ての組合せにおいてについて、個々にミクロ的に比較し、これをミクロ的な類似率として数値化し、その平均や分散を求めることにより、２つの技術文献群間の比較数値としていた（例えば、"パテントマップガイダンス"、特許庁、平成１４年８月４日検索参照、インターネット<http://www5.ipdl.jpo.go.jp/pmgs1/pmgs1/pmgs>、以下非特許文献１という。）。

特開２０００－３４８０１５号公報に記載の知的財産評価装置等には、出願中あるいは登録後の発明等に関する知的財産の財産的価値を評価する知的財産評価装置において、実施利益に関するデータを入力する実施利益入力手段と、各年ごとの複利現価率に関するデータを入力する複利現価率入力手段と、前記実施利益入力手段により入力された各年目ごとの複利現価率に関するデータとを乗算して、各年目ごとの補償金年額の複利現価率を算出する複利現価算出手段と、前記複利現価算出手段により算出された各年ごとの補償金年額の複利現価各年ごとに合算することにより知的財産価額を算出する知的財産価格算出手段と、前記知的財産価格算出手段により算出された知的財産価額を出力する出力手段とを備えた知的財産評価装置、知的財産評価方法等が記載されている。

該知的財産評価装置等では、登録された特許と、それに関連する売上高や利益などを減価償却して現在残存している特許の資産的価値を把握しようとするものである。なお、当該発明では、各特許の価値の評価は、自社評価や他社評価などをランク付けして入力したものを寄与度として評価し、具体的な実施権設定がされていない知的財産の資産価値を算出することが可能であるとされている。

特開２００１－７６０４２号公報に記載のシステム等では、所定の更新間隔を有する第１のデータと更新間隔が当該第１のデータより短い第２のデータとから、経時的に変動しうる評価項目を評価するシステムであって、（ａ）サンプル対象の第１のデータの入力に応じて、第１の評価モデルを作成する手段と、（ｂ）前記サンプル対象の第１のデータを第１の評価モデルに適用し、第１の評価出力を算出する手段と、（ｃ）サンプル対象の第２のデータと第１の評価出力の入力に応じて、第２の評価モデルを作成する手段と、（ｄ）評価対象の第１のデータの入力に応じて、当該第１のデータを第１の評価モデルに適用し、第２の評価出力を算出する手段と、（ｅ）前記評価対象の第２のデータと前記第２の評価出力を第２の評価モデルに適用し、当該評価対象の評価出力を算

出する手段とを備えた経時的に変動しうる評価対象の評価項目を評価するためのシステム、方法および記録媒体が知られている。

該システムでは、１年単位や四半期単位毎に更新される貸借対照表や損益計算書等からの財務データ等のように、更新間隔が比較的長い第１のデータから算出される格付けデータや倒産確率などの企業評価を行うモデル（静的モデル）と、日々変動する株価や金利、為替等のように更新間隔が比較的短い第２のデータとの入力に基づき、その後の変化を予測して動的に企業評価を行うモデル（動的モデル）の２つの評価モデルについて、評価対象の企業データを適用することにより、適時、最新の企業評価を算出することが可能であるとされている。

また、特開平８－２８７０８１号公報、特開２００１－３３７９９２号公報、特開平１０－７４２０５号公報、特開平８－２７８９８２号公報、特開平１１－７３４１５号公報、及び特開２００１－３３１５２７号公報では、ある文書や文章と類似する内容の文書や文章を検索する際に、文書や文章同士の類似度や信憑性が高く、高精度に類似文書を検索することが可能な類似文書検索装置や類似検索システム等が紹介されている。

発明の開示

ところが、非特許文献１に記載のパテントマップや特開平８－２８７０８１号公報、特開２００１－３３７９９２号公報、特開平１０－７４２０５号公報、特開平８－２７８９８２号公報、特開平１１－７３４１５号公報、及び特開２００１－３３１５２７号公報に記載の発明では、例えば、Ａ社に関連した技術文献Ａ群と、Ｂ社に関連した技術文献Ｂ群との間で、技術文献に記載されている内容をマクロ的に比較する要求があったとしても、従来は技術文献Ａ群と技術文献Ｂ群に所属する個々の技術文献どうしを個々ミクロ的に比較し、その複数の演算結果からマクロ的な技術文献群間の比較を導き出していたため、作業効率が悪いという不具合を生じていた。

また、非特許文献１に記載のパテントマップでは、技術比較において同一又は類似の研究開発テーマの内容ごとに、全体的な動向や分布を知ることが可能であるとされているが、企業間において、企業全体の総技術文献を母集団とした各技術の相対的評価を算出することができないため、無形資産の価値評価手法として、定量的、定性的な結果を得られず、信託や投資の評価の対象、企業の特許戦略の決め手となる技術評価の指標を算出することができないという不具合を生じていた。

また、このミクロ的な類似率を平均する計算方法を用いると、例えば、図１９に示す場合において、技術文献群Ａと技術文献Ｂ群とが全く異なる場合には、類似率は０と算出される。また、全ての組合せで求めた平均の類似率も０となるので問題ないように見える。

ところが、第１の技術文献群と第２の技術文献群とがまったく同一の場合であっても、第１の技術文献群に含まれる技術文献A1に対して第２の技術文献群に含まれる技術文献献B1, B2, B3, B4のミクロ的な類似率を求めると、２つの技術文献が全く同一の場合（A1=B1など）にはA1とB1との類似率は１と算出されるが、それ以外の場合には一般に類似率が１になることはない。更にA1以外の、A2、A3、A4などに対する総ての組合せで求めた平均類似率は、１とそれ以下の数値の平均となるので、やはり類似率が１と算出されることはないという不具合を生じる。

また、技術文献の総数が数万件以上となる場合のように、多量の技術文献どうしについて類似率を算出する際には、全ての技術文献の組合せについて類似率を計算する必要があるので、類似率を算出するにあたっての計算量が膨大となるために、計算時間が多く必要となり、類似率の計算結果を素早く表示することができないという不具合を生じている。

また従来のように類似率を算出するにあたり、調査対象と母集団の技術文献をキーワードで切り分け、個々のキーワードが含まれる技術文献の数量と、技術文献の総数との比率を演算し、キーワード総てについて演算した比率を平均して類似率を算出する方法では、キーワードの重要性に応じた重み付けを行なわないと、算出される類似率と実際の感覚的

な類似率との差が大きく開いてしまうという不具合を生じている。

この重み付けをしたキーワードを用いて類似率を算出する際に、全キーワードについてオペレータが重み付けを行なってシソーラス辞書を作成し、その重み付けに基づいて類似率を算出することが可能である。これは理論的に可能ではあるが、実際に膨大な量のキーワードのそれぞれに重みを付けることは結構大変な作業（至難の業）であるし、処理の自動化にはそぐわない。また、個々の技術文献毎に類似率が算出されることには変わりないので、結局は技術文献同士をミクロ的に比較しているにすぎないという不具合を生じていた。

また、非特許文献１に記載のパテントマップでは、パテントマップ作成支援ソフトの価格が、約１５万～５０万円程度であり、その操作には、コンピュータだけでなく特許請求の範囲、図面等を読み取る等高度な技術力と知識力を必要とする。特許調査機関で依頼する場合でも、１件あたり３０万円以上の費用が必要であるとともに、約１ヶ月以上作成時間が必要となる。

従って、資本金や開発費の少ないベンチャー企業等が利用する場合、あるいは出願を急ぐ場合には、パテントマップの利用が制限されることが想定される。

また、従来の知的財産評価装置等では、製品等の研究開発の着手前に過去から最近の情報を広く収集して、競合他社の技術動向の分析や、技術レベルを把握する技術動向調査等の調査を行いにくいという不具合を生じていた。

近年、企業価値に占める無形資産（インタンジブル・アセット）の割合が大きくなるにつれ、無形資産の価値が企業価値を大きく左右するようになってきた。

従って、信託会社は信託の対象に、投資家は投資の対象に、企業は知的財産から産出される利益を重視すべく特許戦略の動向の対象に、それぞれ無形財産を指標として用いる傾向にある。

しかし従来は、投資の参考にするために、技術文献一般を用いて企業

の保有する無形財産を比較するための適切な指標が存在していなかった。

特に、生き残りをかけた企業経営において、新規事業参入や新製品の開発に着手する前段階において開発費を充てる価値のある技術分野なのか、特許出願すべき価値があるのか、出願審査の請求をすべきか否か、権利化の可能性があるのか、ライセンス交渉をした場合の方が利益率が高いか否か等の特許戦略を検討するための指標の存在が、非常に重要となってきている。

そこで本発明は、上記従来の状況に鑑み、企業間において、特許公報等に限られない広範な技術文献群同士を比較し、人の感覚と一致する適当な類似率を算出することによって、定量的、定性的かつ相対的な無形資産の価値を評価することが可能な指標を算出するための類似率算出装置、類似率算出プログラム並びに類似率算出方法を提供することを目的としている。

また本発明は、第１の技術文献群と第２の技術文献群とが全く違ったときだけは類似率が０と算出されるとともに、第１の技術文献群と第２の技術文献群とが同一のときだけは類似率が１と算出されうるものであって、大量で時間のかかる計算を必要とせず、分析者の恣意が混入することによって算出される類似率の値が変わる可能性が少なく、第１の技術文献群と第２の技術文献群との間でマクロ的な類似性の比較結果を算出することが可能な類似率算出装置、類似率算出プログラム並びに類似率算出方法を提供することを目的としている。

また本発明では、比較する技術文献の総数が数万件以上となる場合であっても、比較的短い計算時間で類似率を算出することが可能な類似率算出装置、類似率算出プログラム並びに類似率算出方法を提供することを目的としている。

また本発明では、技術文献群同士をマクロ的に比較することが可能な類似率算出装置、類似率算出プログラム並びに類似率算出方法を提供することを目的としている。

また本発明では、無体財産により企業価値を見極める要求のある投資家等や一般の実務者にも容易に扱うことが可能な類似率算出装置、類似率算出プログラム並びに類似率算出方法を提供することを目的としている。

上記課題を解決するために本発明は、特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との技術的な類似性を判断するための指標を算出する類似率算出装置であって、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数との比を類似率として算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、各混在クラスタに含まれる技術文献の量に応じた値を取る第１の補正値と各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値を取る第２の補正値とを乗算したものを各混在クラスタについて総和を算出し、前記算出し

た全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、個々のクラスタ内の技術文献数のα乗（但し、０＜α）に比例した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、個々のクラスタ内の技術文献数のα乗（但し、０＜α）を、全クラスタ内の技術文献数の平均値等の規格化因子で除算した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第1の技術文献群及び第2の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第1の技術文献群及び第2の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、前記算出した全

クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。また本発明は、前記規格化因子を、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率の最大値のγ乗（但し、$0<\gamma$）としたことを特徴としている。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、第１の技術文献群に含まれる技術文献数Ｍと第２の技術文献群に含まれる技術文献数Ｎとの構成比、Ｎ／Ｍと、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群の技術文献数ｍと第２の技術文献群の技術文献数ｎの混在比、ｎ／ｍとについて、更に構成比と混在比との比を取ったもののτ乗（但し、$0<\tau$）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ

分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、前記期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を任意定数ξ（但し、１＜ξ）の負の指数とした補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

また上記課題を解決するために本発明は、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、前記期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を混在クラスタに含まれる技術文献数で除算したものを任意定数ξ（但し、１＜ξ）の負の指数とした補正値とし、これを各混在クラスタについて総和を算出し、更に前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、前記算出した類似率を記録

手段、表示手段、又は通信手段に出力する出力手段とを備えたことを特徴とする。

本発明によれば、特許文献又は技報等の技術文献から構成される第1の技術文献群と第2の技術文献群との技術的な類似性を判断するための指標を算出する類似率算出装置であって、比較対象となる第1の技術文献群及び第2の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第1の技術文献群及び第2の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第1の技術文献群及び第2の技術文献群の双方の技術文献を含む混在クラスタ数との比を類似率として算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたので、その分解した全クラスタ数と混在クラスタ数の比に基づいて、技術文献群に記載されている技術内容の類似性を示す指標を簡便に算出することが可能となる。

また本発明によれば、類似率算出手段に各混在クラスタに含まれる技術文献の量に応じた値を取る第1の補正値と、各混在クラスタに含まれる第1の技術文献群の技術文献と第2の技術文献群の技術文献との混ざり具合に応じた値を取る第2の補正値とを乗算したものを、各混在クラスタについて総和を算出して、全クラスタ数で除算して類似率を算出する機能を設けたので、補正項1の存在により混在クラスタに含まれる技術文献の量に応じて重要度が高いことを意味付ける補正が可能となるとともに、補正項2の存在により混在クラスタに含まれる技術文献の割合が所定の量に近い程、重要なクラスタであるとして、類似率が高い値を示すように重い重み付けをして、類似率の算出結果を、より人の感覚に合うように補正することが可能となる。

従って、補正項1及び補正項2を用いて類似率を算出することによって、技術文献数量の多い混在クラスタを重要視して類似率を補正すると

ともに、技術文献の混ざり具合が不均一な場合には、類似率を小さい値に補正することが可能となる。

また本発明によれば、類似率算出手段に個々のクラスタ内の技術文献数のα乗（但し、0＜α）に比例した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する機能を設けたので、クラスタ内の技術文献数が多いほど重要なクラスタであるとするような類似率を算出することが可能となる。

また本発明によれば、類似率算出手段に個々のクラスタ内の技術文献数のα乗（但し、0＜α）を、全クラスタ数等の規格化因子で除算して類似率を算出する機能を備えたので、0≦類似率≦1 を保証することが可能となる。また、規格化因子として全クラスタ内の技術文献数の平均値を配置したので、全クラスタ内の技術文献数の平均値を基準として技術文献の量の多少を算出することが可能となる。

また本発明によれば類似率算出手段に、第1の技術文献群の中からm個、第2の技術文献群の中からn個の技術文献を取り出す確率のγ乗（但し、0＜γ）に比例した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する機能を設けた。すなわち、類似率算出手段に（A群の中からm個、B群の中からn個の技術文献を取り出す組合せの数）／（A群とB群とを混ぜ合わせた中からm＋n個の技術文献を取り出す組合せ数）を分子に配置した演算を行なう機能を備えたので、混在クラスタに含まれるA群及びB群の技術文献数の偏り（作為性）に応じて、偏り大の場合は小さい補正値に、偏り小の場合は大きい補正値に類似率を補正することが可能となる。また、規格化因子として、第1の技術文献群の中からm個、第2の技術文献群の中からn個の技術文献を取り出す確率の最大値のγ乗（但し、0＜γ）を配置したので、類似率の算出範囲として0≦類似率≦1 を保証することが可能となる。

また本発明によれば類似率算出手段に、第1の技術文献群に含まれる技術文献数Mと第2の技術文献群に含まれる技術文献数Nとの構成比、

N／Mと、クラスタ分解した結果得られた混在クラスタに含まれる第1の技術文献群の技術文献数mと第2の技術文献群の技術文献数nの混在比、n／mとについて、更に構成比と混在比との比を取ったもののζ乗（但し、0＜ζ）に比例した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する機能を備えたので、A群とB群の技術文献数量の構成比と各クラスタ内における技術文献同士の混在比が同じであるほど類似率を高く算出する（1に近づける）ことが可能となる。

また、構成比と混在比との比の指数ζを　ζ＞1　に設定することによって、A群とB群の技術文献数量の比と、各クラスタ内における技術文献同士の混在比との比が小さい混在クラスタの影響を、類似率の算出結果に大きく反映させないようにすることが可能となる。

また、指数ζを　ζ＝1　に設定することによって、単純にA群とB群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比に応じて類似率を増減させることが可能となる。

また、分子の指数を0＜ζ＜1　に設定することによって、A群とB群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比が大きい場合に類似率の算出結果に対する影響を少なくすることが可能となる。

また本発明によれば類似率算出手段に、第1の技術文献群と第2の技術文献群とを混合した技術文献群の中から第1の技術文献群の技術文献を取り出す確率に前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第1の技術文献群の技術文献を取り出す期待値を算出し、前記期待値と混合クラスタに含まれる第1の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を任意定数ξ（但し、1＜ξ）の負の指数とした補正値を、各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率と算出するようにしたので、ξの値の設定に応じて期待値差に対する類似率の算出結果を敏感に反応させる補正を行なうことが可能となる。

また本発明によれば類似率算出手段に、第1の技術文献群と第2の技術文献群とを混合した技術文献群の中から第１の技術文献群の技術文献を取り出す確率に前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、前記期待値と混合クラスタに含まれる第1の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を混在クラスタに含まれる技術文献数で除算したものを、任意定数ξ（但し、１＜ξ）の負の指数とした補正値とし、これを各混在クラスタについて総和を算出し、更に全クラスタ数で除算して類似率と算出するようにしたので、ξの値の設定に応じて期待値差に対する類似率の算出結果を敏感に反応させる補正を行なうことが可能となる。

図面の簡単な説明

図１は、本発明に係る類似率算出システムの全体構成図である。

図２は、本発明に係る類似率算出装置のブロック図である。

図３は、技術文献Ａ群と技術文献Ｂ群に含まれる技術文献の構成を示す図である。

図４は、類似率の表示処理を示すフローチャートである。

図５は、類似率算出のための入力画面の表示例を示す図である。

図６は、算出した類似率を利用者に通知する類似率表示画面の表示例を示す図である。

図７は、本発明に係る類似率算出装置を用いて技術文献群をクラスタ分解した後の各クラスタの構成を示す図である。

図８は、類似率の算出処理を示すフローチャートである。

図９は、類似率の計算に用いる設定条件を示す図表である。

図１０は、混在クラスタ１には技術文献が多く含まれている状況を表す図である。

図１１は、補正項１（１）を採用した場合の類似率算出例の図表である。

図１２は、補正項２（１）を採用した場合の類似率算出例の図表である。

図１３は、補正項１（１）及び補正項２（１）の双方を採用した場合の類似率算出例の図表である。

図１４は、補正項２（２）を採用した場合の類似率算出例の図表である。

図１５は、補正項１（１）及び補正項２（２）を採用した場合の類似率算出例の図表である。

図１６は、（式３１）に条件１～４を代入した場合の期待値差の算出例を示す図表である。

図１７は、ξ＝１０とした場合において、（式３２）に条件１～４を代入した場合の類似率算出例の図表である。

図１８は、補正項１（１）及び補正項２（３）を採用した場合の類似率算出例の図表である。

図１９は、技術文献Ａ群に含まれる技術文献と、技術文献Ｂ群に含まれる技術文献を個々にミクロ的に比較する従来の状況を示す図である。

発明を実施するための最良の形態

図１は、本発明に係る類似率算出システムの全体構成図である。

同図に示すように、本発明に係る類似率算出システムは、技術文献データベース２０から通信網１０を介して、類似率の算出に必要な技術文献を読み出して、類似率を算出して表示する類似率算出装置３０と、通信網１０を介して各社の技報や、出願済みの特許公報、実用新案公報等の特許文献を含む技術文献を記録する技術文献データベース２０とが設けられている。

通信網１０は、インターネット等の通信網であって、類似率算出装置３０が通信網１０を介して技術文献データベース２０から特許文献等の技術文献に関する情報を取得することが可能となっている。

類似率算出装置３０は、利用者から比較対象の技術文献群に関する情報や、文献どうしの比較条件を入力し、技術文献データベース２０から通信網１０を介して、類似率の算出に必要な技術文献を読み出して、類似率を算出して表示することが可能となっている。

図２は、本発明に係る類似率算出装置のブロック図である。

同図に示すように類似率算出装置３０の情報送受信部には、公衆回線又は通信ネットワーク等の通信網３６４を介して、技術文献データベース２０等の他の通信機器と情報の送受信を行なうことが可能な送受信手段３６５（技術文献群入力手段、技術情報入力手段、又は出力手段の機能を含むものであってもよい）が設けられている。

送受信手段３６５は、技術文献データベース２０から通信網１０を介して、類似率の算出に必要な技術文献を取得することが可能となっている。

また類似率算出装置３０には、利用者から比較対象の技術文献群に関する情報や、文献同士の比較条件を入力するキーボード、マウス等の入力手段３７０（技術情報入力手段の機能を含むものであってもよい）が設けられている。

また類似率算出装置３０には、入力手段３７０を介して入力した各種情報を読み取って後述する情報処理手段３８０に伝達したり、情報処理手段３８０からの指示に基づいてＬＥＤ等に表示指令を出力する入力インターフェース３７１（技術情報入力手段の機能を含むものであってもよい）と、画像や文字等の情報を表示する表示手段３７２（出力手段の機能を含むものであってもよい）と、情報処理手段３８０の指令に基づいて表示手段３７２に対して表示用の画像信号を出力する表示インターフェース３７３（出力手段の機能を含むものであってもよい）とが設けられている。なお、入力手段３７０は、キーボードやマウスに限らず、タブレット等の入力装置を含むものである。

また類似率算出装置３０には、記録媒体３７７を着脱可能に装着する記録媒体装着部３７８と、記録媒体３７７に対して各種情報を記録したり読み出したりする記録媒体インターフェース３７９（技術文献群入力手段、技術情報入力手段、又は出力手段の機能を含むものであってもよい）とが設けられている。なお、記録媒体３７７は、メモリーカード等の半導体や、ＭＯ、磁気ディスク等に代表される磁気記録式、光記録式

等の着脱可能な記録媒体である。

また、類似率算出装置３０には、類似率算出装置３０の全体の制御を行う情報処理手段３８０と、情報処理手段３８０にて実行されるプログラムや各種定数が記録されているＲＯＭや情報処理手段３８０が処理を実行する際の作業領域となる記録手段であるＲＡＭとから構成されるメモリ３８１とが設けられている。

また、情報処理手段３８０（クラスタ分解手段、又は類似率算出手段）は、利用者から比較対象の技術文献群に関する情報や、文献同士の比較条件を入力し、技術文献データベース２０から類似率の算出に必要な技術文献を取得し、記録手段３８４に記録されている類似率の演算プログラムや類似率の算出処理プログラムに基づいて、技術文献同士の類似率を算出する機能を実現することが可能となっている。また、類似率の算出結果を表示手段３７２に表示する機能を実現することが可能となっている。

なお、情報処理手段３８０（クラスタ分解手段）は、文書内の請求項、発明の詳細な説明、図面の簡単な説明、要約などに含まれる言葉（単語、熟語、名詞、動詞、助動詞、形容詞、副詞、助詞など）からなる文を分かち書きしたり、１字、２字など機械的に切り出して技術文献を検索し、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能を実現することが可能となっている。

また情報処理手段３８０（クラスタ分解手段）は、書誌事項などに含まれる項目（ＩＰＣ等の分類、出願日、出願番号、出願人名、発明者、審査請求の有無、補正の有無、国内優先の有無、外国出願の有無、拒絶理由の有無、登録日、登録番号、など）を用いてクラスタ分解する機能を実現することが可能となっている。

また情報処理手段３８０（類似率算出手段）は、クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数との比を算出するなどして、技術文献群同士の類似率を算出する機能を実現することが可能となってい

る。

これらの全ての処理を情報処理手段３８０が実行する代わりに、複数の処理装置に分担して実行するようにしても本発明の目的を達成することが可能である。

また、類似率算出装置３０には、類似率算出装置３０の処理に関する各種定数やネットワーク上の通信機器に通信接続する際の属性情報、ＵＲＬ（Uniform Resource Locators）、ゲートウェイ情報、ＤＮＳ（Domain Name System）等の接続情報、企業の経営に関する情報、特許に関する情報、特許文献、技報、キーワード、技術情報等の各種情報を記録することが可能なハードディスク等の記録手段３８４と、記録手段３８４に記録されている情報を読み出したり記録手段３８４に対して情報を書き込む処理を行う記録手段インターフェース３８５（技術文献群入力手段、技術情報入力手段、又は出力手段の機能を含むものであってもよい）と、時刻を刻むカレンダ時計３９０とが設けられている。

類似率算出装置３０内の情報処理手段３８０と、表示インターフェース３７３、メモリ３８１、記録手段インターフェース３８５、カレンダ時計３９０等を含む各周辺回路はバス３９９で接続されており、情報処理手段３８０にて実行される処理プログラムに基づいて各々の周辺回路を制御する機能を実現することが可能となっている。

前記送受信手段３６５、記録媒体インターフェース３７９、記録手段インターフェース３８５等の技術情報入力手段は、比較対象となる第１の技術文献群及び第２の技術文献群を入力することが可能となっている。

前記送受信手段３６５、入力手段３７０、入力インターフェース３７１、記録媒体インターフェース３７９、記録手段インターフェース３８５等の技術情報入力手段は、キーワードやＩＰＣなどの技術情報を入力することが可能となっている。

前記送受信手段３６５、表示インターフェース３７３、記録手段インターフェース３８５、記録媒体インターフェース３７９、プリンタイ

ンターフェース等の出力手段は、類似率算出手段が算出した類似率を、記録手段、表示手段、又は通信手段に出力することが可能となっている。

図１に示したデータベース２０は、記録手段３８４に記憶されている場合や、ＣＤ－ＲＯＭ、ＣＤ－ＲＷ、ＤＶＤ、ＭＯ等の記憶媒体３７７で提供される場合、通信網３６４を介して他の通信機器から取得する場合も考えられる。

また、上記の類似率算出装置３０は、パーソナルコンピュータ、ワークステーションなど様々なコンピュータを利用して実現することができる。さらに、コンピュータをネットワークで接続して機能を分散して実施するようにしても良い。

本発明に係る類似率算出装置並びに類似率算出プログラムによって算出される技術文献の類似率とは、第１の技術文献群（技術文献Ａ群）と別の第２の技術文献群（技術文献Ｂ群）とを所定のキーワードやＩＰＣなどに基づいてマクロ的に比較した際に算出される数値であって、技術文献群同士がどの程度技術的に関連があるかを示す指標とするための数値をいう。

そして、第１の技術文献群（技術文献Ａ群）と第２の技術文献群（技術文献Ｂ群）は、何らかの属性を持つ技術文献の集まりとする。

本発明では、Ａ社が出願した特許公報やＡ社が発行した技報などの第１の技術文献群（技術文献Ａ群）に記載された技術内容と、Ｂ社が出願した特許公報やＢ社が発行した技報などの第２の技術文献群（技術文献Ｂ群）に記載された技術内容とが、どれだけ類似しているかを定性的に評価する指標としての数値を算出することによって、技術文献同士を容易に比較することが可能となっている。

以下に説明する実施例では、第１の技術文献群（技術文献Ａ群）と第２の技術文献群（技術文献Ｂ群）に記載された技術内容が類似しているほど、類似率は大きい値をとるものと定義している。

なお本発明では、類似率を算出する際に異なる条件を設定した場合であっても、第１の技術文献群（技術文献Ａ群）と第２の技術文献群（技

術文献Ｂ群）との間で算出した類似率と、第３の技術文献群（技術文献Ｃ群）と第４の技術文献群（技術文献Ｄ群）との間で算出した類似率とを直接比較することが可能であるように、類似率が取り得る範囲として、０≦類似率≦１となるような演算を行なうことにしているが、類似率の取り得る範囲はこの範囲に限定されるものではない。

図３は、技術文献Ａ群と技術文献Ｂ群に含まれる技術文献の構成を示す図である。

同図に示すように、技術文献Ａ群は、Ａ１，Ａ２，Ａ３，…ＡＭのＭ個の技術文献から構成されており、技術文献Ｂ群は、Ｂ１，Ｂ２，Ｂ３，…ＢＮのＮ個の技術文献から構成されている。

図４は、類似率の表示処理を示すフローチャートである。

同図に示すように、利用者が技術文献群どうしを比較して、技術内容が類似する度合いを調査する場合には、Ｓ１０「類似率算出指示入力」（以下Ｓ１０のように省略して記載する。）において、類似率算出指示を類似率算出装置３０のキーボード、マウス等の入力手段３７０を操作して入力し、以降の処理を実施させる。

類似率算出装置３０が、Ｓ１００「入力画面読出・表示」にて、類似率算出指示に基づいて、類似率算出に関する各種条件の入力画面の表示情報を記録手段３８４から読み出して、その表示情報に基づいた類似率算出に必要な条件の入力画面を表示手段３７２に表示する。

図５は、類似率算出のための入力画面の表示例を示す図である。

同図に示すように入力画面には、比較対象となっている第１の技術文献群と第２の技術文献群の抽出条件を指定する情報と、キーワードやＩＰＣなどの技術情報を指定する旨の情報が表示されている。利用者は、表示画面に基づいて諸事項を入力することが可能となっている。

クラスタ分解の条件を入力する部分では、特許公報、技報等の対象文献の指定や、全文、請求項部分のみ等の対象部分の設定や、ＩＰＣ、キーワード等のクラスタ分解の尺度等の各種条件を入力することが可能となっている。更に技術文献群の抽出条件として、特許公報の出願日の

期間、業界名称、出典元の企業名、個人名等を入力する項目が表示されている。利用者は、図５に示した入力画面に基づいて、容易に類似率の算出条件を入力したり、予め設けられている複数の算出条件の中から所望の算出条件を選択することが可能となっている。

また図５には、混在クラスタ比を類似率の算出用途に応じて補正するための、補正方法を入力する部分が設けられている。

例えば補正項１として、各混在クラスタに含まれる技術文献の量に応じた値に基づいて、類似率を補正するか否かの補正条件を、利用者が入力することが可能となっている。

また補正項２として、各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値に基づいて、類似率を補正するか否かの補正条件を、利用者が入力することが可能となっている。

なお本発明では、この技術文献との混ざり具合に応じた補正方法として、第１の技術文献群の中からｍ個、第2の技術文献群の中からn個の技術文献を取り出す確率のγ乗（但し、$0 < \gamma$）に比例した補正値を各混在クラスタについて総和を算出し、これを全クラスタ数で除算して類似率の補正等を行なう、「技術文献数の確率」に応じた補正方法を選択することが可能となっている。

また本発明では、第１の技術文献群に含まれる技術文献数Ｍと第２の技術文献群に含まれる技術文献数Ｎとの構成比Ｎ／Ｍと、クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群の技術文献数ｍと第２の技術文献群の技術文献数ｎの混在比、ｎ／ｍとについて、更に構成比と混在比との比を取ったもののζ乗（但し、$0 < \zeta$）に比例した補正値を、各混在クラスタについて総和を算出し、これを全クラスタ数で除算して類似率の補正等を行なう、「技術文献の混在比」に応じた補正方法を選択することが可能となっている。

また本発明では、第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、

前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を任意定数ξ（但し、１＜ξ）の負の指数とした補正値を、各混在クラスタについて総和を算出し、これを全クラスタ数で乗算して類似率の補正等を行なう。「技術文献の期待値差」に応じた補正方法を選択することが可能となっている。

図４に示すＳ１２「類似率算出条件入力」にて、利用者は、表示手段に表示されている案内に基づいて、特許文献、技報、社報、技術論文等の技術文献種別や、比較する技術文献群の指定、クラスタ分解を実施する際に技術文献群から技術文献を抽出する条件となるＩＰＣ又はキーワード等、更に類似率を算出する際の目的に応じた補正情報を、入力手段３７０を介して入力する。

Ｓ１０２「技術文献取得」にて、情報処理手段３８０は、利用者から入力した技術文献種別（例えば特許文献）に基づいて検索するデータベースを特定し、利用者から入力した技術文献群（例えばＡ社の技術文献Ａ群及びＢ社の技術文献Ｂ群）の指定に基づいた技術文献群の取得情報を特定のデータベースに出力する。

Ｓ１３０「技術文献読出」にて、技術文献データベース２０は、類似率算出手段３０から取得した技術文献種別と、技術文献群等に基づいて、データベース内を検索して技術文献を読出して、類似率算出装置３０に送信する。

Ｓ１０４「類似率算出処理」にて、類似率算出装置３０は、データベース２０から取得した技術文献群（例えばＡ社の技術文献Ａ群及びＢ社の技術文献Ｂ群）の中から、利用者指定のＩＰＣやキーワードを共通して含む技術文献を選び出して、クラスタ毎に分解する処理を行なう。

クラスタ分解した結果、技術文献Ａ群に属する技術文献及び技術文献Ｂ群に属する技術文献とが混在しているクラスタを混在クラスタと定義する。本発明では、全クラスタのうち、混在クラスタが存在する割合

に基づいて類似率を算出する。

また、類似率の用途に応じて、混在クラスタに含まれる技術文献の数量や混在確率、混在比率、又はこれらの組合せに応じた補正を行なうことも可能である。

Ｓ１０６「類似率表示処理」にて、類似率算出装置３０は、算出した類似率を表示手段３７２に表示して、利用者に通知する。なお、Ｓ１０６にて類似率を表示手段３７２に表示する代わりに、算出した類似率を送受信手段３６５と通信網１０を介して他の通信機器に送信出力するようにしてもよいし、記録手段インターフェース３８５を介して記録手段３８４に記録出力するようにしてもよいし、記録媒体インターフェース３７９を介して記録媒体３７７に記録出力するようにしてもよい。また、算出した類似率を、印刷用のプリンタインターフェース（図示せず）を介して印刷手段に出力するようにしてもよい。

図６は、類似率算出装置３０が算出した類似率を利用者に通知する類似率表示画面の表示例を示す図である。

同図に示すように、類似率表示画面には、利用者が入力した技術文献群を抽出指定する情報と、キーワードやＩＰＣなどの技術情報をクラスタ分解した際の尺度や、補正方法等の入力情報が確認のために表示されている。

また類似率表示画面には、補正項３として、例えばクラスタ分解した際の所定の特許分類やキーワードに注目して恣意的な重み付けを行なうための補正条件を、各クラスタ毎に利用者が入力することが可能となっている。同図に示す例では、補正項３の数値として「１．０００」を設定している。

また類似率表示画面には、類似率の算出結果と、その類似率を補正するための、α、γ、ζ、ξ等の類似率算出条件を連続的に変更するスライドバーと、各クラスタの補正項を確認するために、分解したクラスタの内容を表示する部分が設けられている。

利用者は、算出された類似率を見ながら、自由に類似率の算出条件を

変更することが可能となっている。利用者がスライドバーを操作した場合には、情報処理手段３８０がカレンダ時計３９０が係数する時間に基づいて、スライドバーの操作完了を判断する。すると、情報処理手段３８０が実施する処理はＳ１０４に分岐して再度類似率を算出し、類似率の演算結果を類似率表示画面に表示する処理を行なう。

図４に示すＳ１４「終了」、Ｓ１０８「終了」及びＳ１４０「終了」にて、類似率算出処理が終了する。

本発明における技術文献のクラスタ分解とは、第１の技術文献群（Ａ群）と第２の技術文献群（Ｂ群）をマクロ的に比較するための「類似率」を算出する際に、キーワードやＩＰＣ等を用いて技術文献を分類することをいう。

本発明を創作するにあたって比較する２つの技術文献群を鳥瞰してみたとき、２つの技術文献群が別々になっていると、非常に計算が複雑になるが、２つを「混ぜて」しまって整理整頓すればずっと計算が容易になるのではないかと、エイヤと「混ぜた」ら案の定類似率の算出に適した様子が見えてきた。双方の技術文献群を混ぜた後、クラスタ分解により分類したところ、一部に両方の技術文献群の構成要素（技術文献）を含むクラスタ（混在クラスタ）が存在し、その分解した全クラスタ数に対する混在クラスタの割合が、我々の通常の感覚としての類似率に近いことがわかった。

先ず、上記のように第１の技術文献群と第２の技術文献群の双方の技術文献を混ぜてひとつの群にする。

混ざった技術文献の群を、何らかの分類法により、ある技術文献の小さな集まり（クラスタという）に分解する。あるクラスタには第１の技術文献群に属する技術文献がｍ個と第２の技術文献群に属する技術文献がｎ個含まれているとする。

技術文献をＩＰＣ（国際特許分類）毎や、技術文献に所定のキーワードが含まれるか否かによって「グループ分け」することを「クラスタ分解」すると定義する。

図７に、本発明に係る類似率算出装置を用いて技術文献群をクラスタ分解した後の各クラスタの構成を示す。

例えば図７に示すように、ＩＰＣ「Ｇ０６Ｆ １７／３０」に分類される技術文献として、第１の技術文献群には「特許文献Ａ１」が、また第２の技術文献群には「特許文献Ｂ１」がそれぞれ存在した場合には、ＩＰＣ「Ｇ０６Ｆ １７／３０」のクラスタには、「特許文献Ａ１」と「特許文献Ｂ１」の要素が含まれる。

また例えば、キーワードとして「テキスト処理」という文言を含む技術文献が、第１の技術文献群には「技術文献Ａ２」が、また第２の技術文献群には「技術文献Ｂ２」及び「技術文献Ｂ３」が存在した場合には、キーワード「テキスト処理」のクラスタには「技術文献Ａ２」と「技術文献Ｂ２」、「技術文献Ｂ３」の要素が含まれる。

なお、クラスタ分解の方法には、技術文献群の個々の技術文献の属性により２通りの扱いがあり、それは以下の通りである。

１．外的な基準がある属性（属性１型と定義する）の場合は、その属性それぞれでクラスタを構成できる。例えば、特許公報等の技術文献でいえば、出願日の日付やＩＰＣなど、一意に決まる技術文献である。

２．内的な関係で属性が決まる値（属性２型と定義する）は、前処理として多変量解析（クラスタ分析）などによるクラスタ化が必要である。例えば特許公報技術文献の中では、要約や請求項などの文書に外的な基準をあてはめることが難しいため、文書間のミクロ的な類似率を別途定義し、それに基づいて多変量分解を行った結果を用いてクラスタを構成する。なお、文書間のミクロ的な類似率については、ＴＦＩＤＦ法など、一般的に広く用いられているものを使用することにより、分析者の恣意の混入を防ぐことが可能である。

情報処理手段３８０等のクラスタ分解手段は、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、技術情報入力手段を介して入力した技術情報を含む技術文献を検索し、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解している。

本発明の実施例では、混在クラスタを以下のように定義する。

図７に示すＩＰＣ「Ｇ０６Ｆ　１７／３０」のクラスタには、技術文献Ａ群に属する「特許文献Ａ１」と、技術文献Ｂ群に属する「特許文献Ｂ１」とが混在している。このように、技術文献Ａ群に属する技術文献、及び技術文献Ｂ群に属する技術文献が混在しているクラスタを混在クラスタと定義する。

本発明の実施例では、非混在クラスタを以下のように定義する。

例えば図７に示すように、ＩＰＣ「Ｂ０１」に分類される技術文献として、技術文献Ａ群には「特許文献Ａ-３」が存在するが、技術文献Ｂ群にはＩＰＣ「Ｂ０１」に分類される技術文献が存在しない場合には、ＩＰＣ「Ｂ０１」のクラスタには「特許文献Ａ３」のみが要素として含まれる。

また図７に示すように、例えばキーワードとして「無機化合物」という文言を含む技術文献は、技術文献Ａ群には存在しないが技術文献Ｂ群には「技術文献Ｂ１」が存在した場合には、キーワード「無機化合物」のクラスタには「技術文献Ｂ１」が要素として含まれる。

このように、技術文献Ａ群に属する技術文献と、技術文献Ｂ群に属する技術文献とが混在していないクラスタを非混在クラスタと定義する。

図８は、類似率の算出処理を示すフローチャートである。

情報処理手段３８０が実施する処理が、図４に示したＳ１０４に進むと、情報処理手段３８０が実施する処理はＳ２００に分岐してきて、Ｓ２００以降の処理を実施する。

類似率算出装置３０の情報処理手段３８０は、Ｓ２００「技術文献Ａ群と技術文献Ｂ群とを混同する」にて、Ｓ１０２「技術文献取得」によってデータベースから取得した技術文献群（例えばＡ社の第１の技術文献群及びＢ社の第２の技術文献群）を混合して、１つの技術文献群にする処理を行なう。

Ｓ２０２「クラスタ分解処理」にて情報処理手段３８０は、キーワードやＩＰＣ等の技術情報に基づいてクラスタ分解処理を行う。次のＳ２

０４「補正項１の計算式を設定」にて、情報処理手段３８０は、混在クラスタに含まれる技術文献の数量に応じて類似率を補正する旨の指示を利用者から入力している場合には、その指示に基づいた補正項の数式を選択する処理を行なう。ここでは、補正の内容に応じて補正項１に所定の数式を代入する処理を行なう。

補正項１は、混在クラスタに含まれる技術文献の量が多い程、重要なクラスタであると考えて類似率が高くなるように重い重み付けをして類似率の補正を行なうための補正項である。

混在クラスタに含まれる技術文献の数量に応じて類似率を補正しない場合には、補正項１＝１（定数）を代入する。

Ｓ２０６「補正項２の計算式を設定」にて情報処理手段３８０は、混在クラスタに含まれる技術文献Ａと技術文献Ｂとの混ざり具合に応じて類似率を補正する旨の指示を利用者から入力している場合には、その指示に基づいた補正項の数式を選択する処理を行なう。ここでは、補正の内容に応じて補正項２に所定の数式を代入する処理を行なう。

補正項２は、混在クラスタに含まれる技術文献の割合が所定の量に近い程、重要なクラスタであると考えて類似率が高くなるように重い重み付けをして類似率の補正を行なうための補正項である。

混在クラスタに含まれる技術文献の混ざり具合に応じて類似率を補正しない場合には、補正項２＝１（定数）を代入する。

Ｓ２０８「補正項３の値を設定」にて情報処理手段３８０は、クラスタ分解した際の所定の特許分類やキーワードに注目して恣意的な重み付けを行なって、類似率を補正する旨の指示を利用者から入力している場合には、その指示に基づいた補正項の数式を選択する処理を行なう。ここでは、補正の内容に応じて補正項３に所定の値を代入する処理を行なう。クラスタ分解した際の所定の特許分類やキーワードについて特に注目しない場合には、補正項３＝１（定数）を代入する。

Ｓ２１０「類似率算出」にて情報処理手段３８０は、各混在クラスタについて補正項１、補正項２、補正項３の各補正項を乗算して総和を算

出する。更に規格化するために全クラスタ数で除算して類似率を算出する処理を行なう。

Ｓ２１２「終了」にて、類似率算出処理のサブルーチンを終了して、元の処理に戻る。

図９に、類似率の計算に用いる設定条件を示す。

図９は比較対象となる第１の技術文献群及び第２の技術文献群と、各群の技術文献を４つのクラスタに分解した場合の各クラスタ１～４に存在する各技術文献数を示す図表である。同図右端に示す「期待する類似率」の値は、技術文献の類似性の判断を行なっている複数の専門家にヒアリングを行なった結果、条件１～４の場合に、算出されることを期待する類似率の値を示したものである。そして、その期待する類似率の値に対して許容され得ると思われる範囲は、同図に示すように許容範囲＝±0.050程度である。

したがって、本発明に係る類似率算出装置を用いて類似率を算出した結果、図９に示す許容範囲内で類似率が算出されれば、技術文献同士の比較が最適に行なわれていることを示している。

基本型１：補正項を考慮しない場合の類似率（基本型1）の算出例

以下に、補正項を用いない基本型の類似率（基本型１）の算出例を示す。この類似率（基本型１）の算出例は、混在クラスタ抽出法により技術文献の類似率を演算するものである。

第１の技術文献群に含まれる技術内容と、第２の技術文献群に含まれる技術内容とが、どれだけ類似しているかの度合（類似率の値の大きさ）は、「混在クラスタの数量」に比例するものと考えられる。

また類似率を、０≦類似率≦１の範囲に設定するために、例えば、「混在クラスタ数」を、「混在クラスタ数と非混在クラスタ数の総和」である「全クラスタ数」で除算した混在クラスタを算出すると、技術文献群同士の類似率として以下の（式1）が得られる。

混在クラスタを考慮した類似率算出方法を混在クラスタ抽出法と定義する。下記に示す（式１）は最も基本的な考え方である。下記の（式

1）では、クラスタ分解した結果得られた全クラスタ数と、第1の技術文献群及び第2の技術文献群の双方の技術文献を含む混在クラスタ数との比（以下混在クラスタ比と呼ぶ）を類似率として算出する一例を示している。したがって、全クラスタ数と混在クラスタ数の比の算出のしかたは、下記の（式1）に限定されるものではない。

$$\text{類似率（基本型1）} = \frac{1}{\text{全クラスタ数}} \sum_{\text{クラスタ}=1}^{\text{全クラスタ数}} \delta$$

$$= \frac{\text{混在クラスタ数}}{\text{全クラスタ数}} \quad \cdots\cdots(\text{式}1)$$

但し、

δ＝混在クラスタの場合……1

　　非混在クラスタの場合…0

先に述べたように、類似率とは第1の技術文献群に記載されている技術内容と第2の技術文献群に記載されている技術内容とがどれだけ類似しているかを示す数値である。

また、混在クラスタ数とは、第1の技術文献群に属する技術文献及び第2の技術文献群に属する技術文献が混在しているクラスタの数を示す数値である。

全クラスタ数とは、第1の技術文献群の技術文献又は第2の技術文献群の技術文献が存在するクラスタの全数を示す数値である。

以下に、類似率（基本型1）の計算式を用いた場合の計算結果について説明する。

第1の技術文献群と第2の技術文献群について、所定のキーワードやＩＰＣ等を用いてクラスタ分解を行なった結果、全クラスタ数が１０個であって、混在クラスタ数が3個であった場合には、類似率（基本型1）＝３／１０＝０．３　と算出される。

また、全クラスタ数が４個であって、混在クラスタ数が２個であった場合には、類似率（基本型１）＝２／４＝０．５と算出される。

第１の技術文献群と第２の技術文献群に含まれる技術文献を、キーワードやＩＰＣ等を用いてクラスタ分解し、その分解した全クラスタ数と混在クラスタ数の比を類似率として算出することによって、技術文献群同士の類似率の基礎部分となる値を算出することが可能となる。

また、類似率を算出する際に、混在クラスタ数を全クラスタ数で除算することによって、算出される類似率の値を、０≦類似率≦１　の範囲に設定することが可能となる。

以下に、類似率（基本型１）を用いた場合の発明の効果について説明する。

第１の技術文献群と第２の技術文献群に含まれるキーワードやＩＰＣ等を用いてクラスタ分解し、その分解した全クラスタ数と混在クラスタ数の比に基づいて類似率を算出することによって、技術文献群同士がどの程度技術的に類似しているかを示す指標を簡便に算出することが可能となる。ここで算出される類似率は、われわれが常識的に考えた技術文献群同士の類似の程度と割合一致していることがわかった。

また本発明では、算出する類似率の値を、０≦類似率≦１　の範囲に設定する演算を行なっているので、全クラスタ数量や混在クラスタの数量、また技術文献群に含まれる技術文献の量の多少に関わらず一定の指標を算出することが可能となる。

更に、より多くの条件下で第１の技術文献群と第２の技術文献群を比較した類似率と、第１の技術文献群と第３の技術文献群とを比較した類似率とを直接対比することも可能となる。

基本型２：補正項を考慮した場合の類似率（基本型２）の算出例

以下に、補正項を考慮した場合の類似率（基本型２）の算出例を示す。この類似率（基本型２）の算出例は、前記類似率（基本型１）の算出例に対して補正項１～３を加味したものとなっている。

上記の（式１）を用いて類似率を算出すると、混在クラスタ数に比例

した類似率が簡単な数式を用いてたいへん素早く算出できるという利点がある。

上記の最も基本的な（式１）は、たとえば多くの技術文献を含むクラスタと少数の技術文献しか含まないクラスタが対等の寄与を持つ結果となることでもわかるように、個々のクラスタ内の技術文献数の大小を考慮していないという欠点があるために、混在クラスタ内に多くの技術文献が含まれる場合であっても、２つしか技術文献が含まれない場合であっても同一の類似率が算出されてしまい、われわれが常識的に考えた類似の程度と異なってしまう場合があるという不具合を生じる可能性がある。

混在クラスタに含まれる技術文献の量の他にも、混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献の混ざり具合（第１の技術文献群の技術文献と第２の技術文献群の技術文献との割合）や、特定の特許分類やキーワードに注目したい場合の恣意的な重み付けなどによって、算出される類似率の値を補正したい場合が生じる。

図１０は、混在クラスタ１に技術文献が多く含まれている状況を示す図である。

図１０に示す例では、クラスタ1（混在クラスタ）には、技術文献が多く含まれているので重要なクラスタであると考えられ、類似率計算の際に最も寄与が大きくなると良い。

別のクラスタ（例えばクラスタ2，クラスタ3，クラスタ4など）は、含まれている技術文献が少ないので重要なクラスタではないと思われるので、クラスタ１の寄与に比べるとずっと小さくなるのが望ましい。

図１０の例のような状況にある場合、クラスタ1に対し、クラスタ2，クラスタ3，クラスタ4の影響を軽視すべき場合がある。なお、含まれる技術文献数量が少ないクラスタの存在を無視しない場合には、算出される類似率の値は0.5まで下がってしまう。

そこで以下の（式２）に示すように、（式1）のδ（クラスタが混在

クラスタである場合には$\delta = 1$とし、それ以外の場合には、$\delta = 0$とする）に対して補正項を乗算することにする。なお、補正によって類似率の範囲が、0≦類似率≦1 の範囲を超えないようにするためには適当な規格化因子が必要である。

$$\text{類似率（基本型２）} = \frac{1}{\text{全クラスタ数}} \times \sum_{\text{クラスター}=1}^{\text{全クラスタ数}} \{(\text{補正項１}) \times (\text{補正項２}) \times (\text{補正項３}) \times \delta\} \quad \cdots (\text{式２})$$

但し、

δ＝混在クラスタの場合……１

　　非混在クラスタの場合…０

（式２）に示す補正項１は、混在クラスタに含まれる技術文献の量に応じて類似率を算出するための補正項である。この補正項１は、混在クラスタに含まれる技術文献の量が多い程、重要なクラスタであると考えて類似率が高くなるように重い重み付けをして類似率の補正を行なう補正項である。

また逆に補正項１は、混在クラスタに含まれる技術文献の量が少ない程、重要なクラスタでないと考えて類似率が低くなるように軽い重み付けをして類似率の補正を行なうことも可能な補正項である。

また補正項１は、各混在クラスタに含まれる技術文献の量に応じた値を取る第１の補正値を算出することが可能な他の計算式を用いた補正項であってもよい。

（式２）に示す補正項２は、混在クラスタに含まれる技術文献Ａと技術文献Ｂの混ざり具合（技術文献Ａと技術文献Ｂの割合）に応じて類似率を算出するための補正項である。

補正項２は、混在クラスタに含まれる技術文献の割合が所定の量に近

い程、重要なクラスタであると考えて類似率が高くなるように重い重み付けをして類似率の補正を行なう補正項である。

また補正項２は、各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値を取る第２の補正値を算出することが可能な補正項である。

（式２）に示すように類似率は、補正項１、補正項２、又は補正項３を全ての混在クラスタについての総和を算出し、該総和を全クラスタ数で除算する演算を行なっている。

補正項２を算出する際に用いる技術文献の「混ざり具合」の意味は、以下のとおりである。

ある混在クラスタに含まれる第１の技術文献群の技術文献、及び第２の技術文献群に含まれる技術文献の混ざり具合に注目して、双方の技術文献がよく混ざっているとき、すなわち双方の技術文献数が偏っていないときに重要なクラスタと考えて重い重みを付け、よく混ざっていない場合、すなわち技術文献数が片方の技術文献群のものに偏っている場合に、重要ではないクラスタと考えて軽い重み付けをするための補正項目である。

言い換えると、たとえばある混在クラスタに含まれる、第１の技術文献群の技術文献と第２の技術文献群の技術文献の数量が、第１の技術文献群と第２の技術文献群から無作為に抽出したときの期待値に近いものは重く、遠いものは軽くする補正項である。

補正項３とは、特定の特許分類やキーワードに注目したい場合に、恣意的な重み付けを行なって類似率を算出するための補正項である。この項は技術文献群同士を比較する者が個別設定する項であるので、今回は考慮せずに定数「1」を代入しておく。

応用型１：補正項１（１）の算出例

$$\text{補正項１（１）}=\frac{(\text{クラスタ内の技術文献数})^{\alpha}}{(\text{規格化因子})}$$

$$=\frac{(\text{クラスタ内の技術文献数})^{\alpha}}{\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}(\text{クラスタ内の技術文献数})^{\alpha}} \quad \cdots(\text{式３})$$

補正項１（１）を考慮した類似率（式４）の算出例を以下に示す。

$$\text{類似率}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\times\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left\{\frac{(\text{クラスタ内の技術文献数})^{\alpha}}{\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}(\text{クラスタ内の技術文献数})^{\alpha}}\times(\text{補正項２})\times(\text{補正項３})\times\delta\right\} \quad \cdots(\text{式４})$$

補正項１（１）では、類似率が混在クラスタに含まれる技術文献の量に応じて大きな値をとるように補正するために、「クラスタ内の技術文献数」のα乗（但し、０＜α）を分子に配置している。そして、類似率の算出範囲として　０≦類似率≦１　を保証するために、補正項１（１）の式では規格化因子を分母に配置している。

（式４）に示す補正項１（１）の演算では、分子に配置したクラスタ内の技術文献数が多い場合であっても類似率の値が１を超えないようにするためと、技術文献の量の多少の判断基準を設けるために、規格化因子として、全クラスタ内の技術文献数の平均値を配置している。なお、規格化因子は、全クラスタ内の技術文献数のα乗の総和を算出し、全クラスタ数で除算した値を配置してもよい。この規格化因子は、０≦類似率≦１　を保証することが可能な項であればよく、（式４）の数式に限定されるものではない。

更に、含まれる技術文献の量が少ない混在クラスタの影響を、類似率

の算出結果に大きく反映させたくない場合には、分子の指数αを　$\alpha > 1$　に設定する。

また、単純にクラスタ内の技術文献数の量に応じて類似率を増減させる要望がある場合には、$\alpha = 1$　に設定する。

また、クラスタに含まれる技術文献の量に応じて類似率を算出するとともに、技術文献が多量に含まれるクラスタの存在による類似率の算出結果の影響を少なくする必要がある場合には、$0 < \alpha < 1$　に設定するとよい。

以下に「応用型１：補正項１（１）」の計算式の分子と分母の構成による作用について説明する。

式４に説明するように「クラスタ内の技術文献数」を補正項１（１）の分子に配置したので、クラスタ内の技術文献数に比例した類似率を算出することが可能となる。

また、「規格化因子」を補正項１（１）の分母に配置したので、$0 \leqq$ 類似率 $\leqq 1$　を保証することが可能となる。そして、補正項１（１）の規格化因子として、全クラスタ内の技術文献数の平均値を配置したので全クラスタ内の技術文献数の平均値を基準として、技術文献の量の多少を算出することが可能となる。

更に、分子の指数αを　$\alpha > 1$　に設定することによって、混在クラスタに含まれる技術文献の量が少ない混在クラスタの影響を、類似率の算出結果に大きく反映させないようにすることが可能となる。また、分子の指数を$\alpha = 1$　に設定することによって、単純にクラスタ内の技術文献数の量に応じて類似率を増減させることが可能となる（単純含数比較）。また、分子の指数を$0 < \alpha < 1$　に設定することによって、技術文献が多量に含まれるクラスタの存在による類似率の算出結果の影響を少なくすることが可能となる。

以下に、「応用型１：補正項１（１）」の計算式（式４）に、図９に示した各条件を代入した場合の計算例を示す。なお、算出結果は、図１１に、補正項１（１）を採用した場合の類似率算出例（補正項１（１）に

条件１～４を代入した場合の計算結果）の図表として示す。

補正項１（１）のみを考慮して他の補正項を考慮しない場合であって（すなわち補正項２＝１、補正項３＝１とする）、単純に混合クラスタ内に含まれる技術文献数の比較を行なう場合（すなわちα＝-1としたとき）に、技術文献群同士を比較する条件として、条件1～4を設定した場合の類似率の試算結果を以降に示す。

下式（式５）に、計算例４－１（式4に条件１を代入した場合）の計算結果について説明する。

条件１の場合には、各混在クラスタ（本実施例の場合には、クラスタ１及びクラスタ２）に含まれる技術文献数は、それぞれ３個である。したがって、クラスタに含まれる技術文献の量による類似率の補正の影響は少ないことが期待される。

$$\text{類似率（式４，条件１）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\left\{\frac{(\text{クラスタ内の技術文献数})^{1}}{\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}(\text{クラスタ内の技術文献数})^{1}}\times1\times1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{3}{(3+3+2+4)/4}+\frac{3}{(3+3+2+4)/4}\right)=0.5 \qquad \text{...(式５)}$$

上記（式５）にて算出した類似率（式4に条件１を代入した場合）＝0.5 の値は、（式１）による類似率の演算結果と一致しており、補正項１（１）を挿入した場合であっても、われわれが常識的に考えた類似率の程度と大きくずれてはいない。また、クラスタ内の技術文献数量がそれぞれ３、３、２、４程度であるので、全てから同じ程度の寄与があるべきで、ここで類似率＝0.5 と算出された結果は、われわれが常識的に考えた類似の程度（約 0.30 程度）から大きく外れてはおらず、おおよそ要件を満足しているものとなっている。

下式（式６）に、計算例４－２（式4に条件２を代入した場合）の計算結果について説明する。

条件２の場合のクラスタ１に含まれる技術文献の量は、クラスタ2～クラスタ４に含まれる技術文献の量よりも際立って多いので、類似率を算出する際には、クラスタ1に含まれる技術文献の量の影響を重視して類似率を大きく算出するべきなのは明らかである。

$$\text{類似率（式４，条件２）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left\{\frac{(\text{クラスタ内の技術文献数})^{1}}{\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}(\text{クラスタ内の技術文献数})}\times 1\times 1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{150}{(150+3+2+4)/4}+\frac{3}{(150+3+2+4)/4}\right)=0.962 \quad \ldots(\text{式６})$$

上記（式６）にて算出した類似率（式４に条件２を代入した場合）＝0.962の値は、クラスタ１に含まれる技術文献の量の多さに引っ張られ、類似率＝0.5（式４に条件１を代入した場合に算出した類似率）から類似率 0.962（式４に条件２を代入した場合に算出した類似率）に補正された。

以下に式６（式４に条件2を代入した場合）の効果について説明する。式６の演算処理によって、クラスタに含まれる技術文献の量が他のクラスタに含まれる技術文献の量よりも多い場合に、その技術文献の量を類似率の算出結果に反映させることが可能となる。これは、クラスタ1が類似率を算出する際の傾向のほぼ全てを代表しているので、このクラスタ１の性質が類似率を決めるように働いているのが見える。

そしてこの類似率の算出結果は、われわれが常識的に考えた類似の程度と、かなり一致していることがわかった。

下式（式）に、計算例４－３（式4に条件を３代入した場合）の計算

結果について説明する。

条件３の場合には、条件２の場合とクラスタに含まれる技術文献の量の総和は同じであるが、クラスタ１に含まれる技術文献の量のみが際立って多い状況ではないので、類似率を算出する際にクラスタ１に含まれる技術文献の量の影響が条件２の場合程は生じないことが望ましい。

$$\text{類似率（式４，条件３）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\left\{\frac{(\text{クラスタ内の技術文献数})^{1}}{\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}(\text{クラスタ内の技術文献数})}\times 1\times 1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{70}{(70+3+82+4)/4}+\frac{3}{(70+3+82+4)/4}\right)=0.459 \quad \text{……(式７)}$$

上記（式７）にて算出した類似率（式４に条件３を代入した場合）0.459 の値は、クラスタ１に含まれる技術文献の量の多さが、他のクラスタ３よりも少し少ない程度であることから、類似率の補正にはほとんど関与しないように補正される。

以下に、（式７）の計算結果（式４に条件３を代入した場合）の効果について説明する。

補正項１（１）の演算処理を行なうことによって、クラスタに含まれる技術文献の量が多い場合であっても、他のクラスタに含まれる技術文献の量と大差が無い場合には、その技術文献の量を類似率の算出結果にあまり反映させないようにすることが可能となる。

この（式７）による類似率の算出結果は、クラスタ１とクラスタ３の影響が大きく出るように補遺性が働いているので、われわれが常識的に考えた類似の程度（約 0.20 程度）と大きくずれてはおらず、ほぼ狙いどおりの値が得られている。

下式（式８）に、計算例４－４（式4に条件を4代入した場合）の計算結果について説明する。

条件４の場合には、条件３の場合とクラスタに含まれる技術文献の量の総和は同じであるが、クラスタ1及びクラスタ2に含まれる第1の技術文献群と第２の技術文献群との割合が極端に不均等である場合である。したがって、混合クラスタに含まれる技術文献数が多いからといって類似率を大きく算出しないことが望ましい。

$$\text{類似率（式４，条件４）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項2})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\left\{\frac{(\text{クラスタ内の技術文献数})^{1}}{\left\{\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}(\text{クラスタ内の技術文献数})^{1}\right\}}\times1\times1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{52}{(52+21+82+4)/4}+\frac{21}{(52+21+82+4)/4}\right)=0.459 \quad \cdots(\text{式８})$$

上記の（式８）にて算出した類似率（式4に条件4を代入した場合）＝0.459の値は、クラスタ１及びクラスタ2に含まれる技術文献の量が多くても、他のクラスタ３よりも少し少ない程度であることから、類似率の補正にはほとんど関与しないように補正される。

以下に、（式８）の計算結果（式4に条件4を代入した場合）の効果について説明する。

（式８）の演算処理によって、クラスタに含まれる技術文献の量が多い場合であっても、他のクラスタに含まれる技術文献の量と大差が無い場合には、その技術文献の量を類似率の算出結果にあまり反映させないようにすることが可能となるが、条件4の場合には類似率は数パーセントの値になることが感覚上望ましい。

この条件４の場合には、補正項1（1）の処理だけでは人の感覚と一致しない部分が生ずる可能性があるために、以降で説明する補正項２が有用となる。但し、クラスタ３、1、2の影響が大きくなっているので、

補正項１（１）の役割は十分に果たしているといえる。また、補正項１（１）の処理を行なうことによって、技術文献数の多いクラスタが存在する場合には、そのクラスタに含まれる技術文献数量の多さを類似率に反映することが可能となっている。

図１１に、補正項１（１）を採用した場合の類似率算出例（補正項１（１）に条件１～４を代入した場合の計算結果）の図表を示す。

応用型２：補正項２（１）の算出例

以下に示す補正項２（１）の計算式（式９）は、混在クラスタ内の技術文献の混在確率に応じて補正を行なうために構成したものである。

$$\text{補正項２（１）}=\frac{(\text{混在確率})^{\gamma}}{(\text{混在確率の最大値})^{\gamma}}$$

$$=\frac{(\text{A群の中からm個、B群の中からn個の技術文献を取り出す確率})^{\gamma}}{(\text{A群の中からx個、B群の中からy個の技術文献を取り出す確率})^{\gamma}}$$

$$=\frac{\left(\dfrac{\text{A群の中からm個、B群の中からn個の技術文献を取り出す組合せ数}}{\text{A群とB群とを混ぜ合わせた中からm＋n個の技術文献を取り出す組合せ数}}\right)^{\gamma}}{\left(\dfrac{\text{A群の中からx個、B群の中からy個の技術文献を取り出す組合せ数}}{\text{A群とB群とを混ぜ合わせた中からm＋n個の技術文献を取り出す組合せ数}}\right)^{\gamma}}$$

$$=\frac{\left(\dfrac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}{MAX\left(\dfrac{{}_{M}C_{x}\times{}_{N}C_{y}}{{}_{M+N}C_{m+n}}\right)^{\gamma}} \qquad \text{...(式9)}$$

但し、

Ｍ：第１の技術文献群（Ａ群）に含まれる技術文献数

Ｎ：第２の技術文献群（Ｂ群）に含まれる技術文献数

ｍ：所定のクラスタに含まれる第１の技術文献群（Ａ群）の技術文献数

ｎ：所定のクラスタに含まれる第２の技術文献群（Ｂ群）の技術文献数

γ：任意定数 $\gamma > 0$

上記補正項２（１）を考慮した類似率（式１０）の算出例を以下に示す。

$$\text{類似率} = \frac{1}{\text{全クラスタ数}} \sum_{\text{クラスタ}=1}^{\text{全クラスタ数}} \{(\text{補正項１}) \times (\text{補正項２}) \times (\text{補正項３}) \times \delta\}$$

$$= \frac{1}{\text{全クラスタ数}} \sum_{\text{クラスタ}=1}^{\text{全クラスタ数}} \left\{ (\text{補正項1}) \times \frac{\left(\frac{{}_{M}C_{m} \times {}_{N}C_{n}}{{}_{M+N}C_{m+n}} \right)^{\gamma}}{\text{MAX}\left(\frac{{}_{M}C_{x} \times {}_{N}C_{y}}{{}_{M+N}C_{m+n}} \right)^{\gamma}} \times (\text{補正項3}) \times \delta \right\} \quad \text{...(式10)}$$

（式１０）の補正項２（１）では、類似率が、混在クラスタに含まれる第１の技術文献群（Ａ群）及び第２の技術文献群（Ｂ群）の技術文献数の確率に応じて大きな値をとるように補正するために、第１の技術文献群（Ａ群）の中からｍ個、第２の技術文献群（Ｂ群）の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）を分子に配置している。

類似率の算出範囲を　０≦類似率≦１　を保証するために、例えば（式１０）に示すように、第１の技術文献群（Ａ群）の中からｍ個、第２の技術文献群（Ｂ群）の中からｎ個の技術文献を取り出す確率の最大値のγ乗（但し、０＜γ）を規格化因子として分母に配置している。

規格化因子は、０≦類似率≦１　を保証することが可能な項であればよく、（式１０）に示した規格化因子に限定されるものではない。

以下に、指数γの設定条件について説明する。

単純に混在クラスタに含まれるＡ群及びＢ群の技術文献数が、Ａ群及びＢ群の技術文献群から無作為に抽出した際の分布に近い度合いに比

例して類似率の値を補正する必要がある場合には、指数γを　$\gamma=1$　に設定するとよい。

また、混在クラスタに含まれるA群及びB群の技術文献数が、A群及びB群の技術文献群から無作為に抽出した際の分布に近いほど重要視して大きな値に補正する必要がある場合、又は、A群及びB群の技術文献群から無作為に抽出した際の分布に遠いほど軽視して小さな値に補正する必要がある場合には、指数γを　$\gamma>1$　に設定するとよい。

また、混在クラスタに含まれるA群及びB群の技術文献数が、A群及びB群の技術文献群から無作為に抽出した際の分布に近くなくても重要視して補正する必要がある場合には、指数γを　$0<\gamma<1$　に設定するとよい。

以下に、応用型2：補正項2（1）の計算式（式10）に，図9に示した各条件を代入した場合の計算例を示す。なお、算出結果は、図12に、補正項2（1）を採用した場合の類似率算出例（補正項2（1）に条件1～4を代入した場合の計算結果）の図表として示す。

補正項2（1）では、（A群の中からm個、B群の中からn個の技術文献を取り出す組合せの数）／（A群とB群とを混ぜ合わせた中からm＋n個の技術文献を取り出す組合せ数）を分子に配置したので、混在クラスタに含まれるA群及びB群の技術文献数の偏り（作為性）に応じて、偏り大の場合は小さい補正値に、偏り小の場合は大きい補正値に類似率を補正することが可能となる。本実施例では、偏りが大きい場合には補正値を小さくして類似率を小さく算出することとし、逆に偏りが小さい場合には補正値を大きくして類似率を大きく算出することとしている。

規格化因子として分母に（A群の中からx個、B群の中からy個の技術文献を取り出す組合せの数）／（A群とB群とを混ぜ合わせた中からm＋n個の技術文献を取り出す組合せ数）を配置したので、x、yは分母を最大にする数の組合せであることから類似率の算出範囲として　$0\leqq$類似率$\leqq 1$　を保証することが可能となる。

更に、分子の指数γを　$\gamma=1$　に設定することによって、単純に混

在クラスタに含まれるＡ群及びＢ群の技術文献数が、Ａ群及びＢ群の技術文献群から無作為に抽出した際の分布に近い度合いに比例して類似率の値を補正することが可能となる。

また、分子の指数γを　γ＞１　に設定することによって、混在クラスタに含まれるＡ群及びＢ群の技術文献数が、Ａ群及びＢ群の技術文献群から無作為に抽出した際の分布に近いほど重要視して大きな値に補正することが可能となる。また、Ａ群及びＢ群の技術文献群から無作為に抽出した際の分布に遠いほど軽視して小さな値に補正することが可能となる。

また、混在クラスタに含まれるＡ群及びＢ群の技術文献数が、Ａ群及びＢ群の技術文献群から無作為に抽出した際の分布に近くなくても重要視して補正する必要がある場合には、分子の指数γを、０＜γ＜１に設定するとよい。

下式（式１１）に、計算例１０－１（式１０に条件１を代入した場合）の計算結果について説明する。

補正項２（１）のみを考慮して他の補正項の作用を考慮しない場合であって（すなわち補正項１＝１、補正項３＝１とする）、単純に混在確率に基づいて比較を行なう場合（すなわちγ＝１とした場合）に、技術文献群同士を比較する条件を、条件１～４に設定したときの類似率の試算結果は、以下のとおりである。

下記の（式１１）に示すように、条件１の場合には、各混在クラスタ１に含まれる技術文献の混在確率は、０．４０９と算出される。また、同様にクラスタ２に含まれる技術文献の混在比率も、０．４０９と算出される。

$$\text{混在確率（条件１，クラスタ１）}=\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{6}C_{2}\times{}_{6}C_{1}}{{}_{6+6}C_{2+1}}\right)=\frac{{}_{6}C_{2}\times{}_{6}C_{1}}{{}_{12}C_{3}}$$

$$=\frac{15\times6}{220}=0.409 \qquad \text{...(式１１)}$$

一方、分母の規格化因子は混在クラスタ１の混在確率の最大値であるので、以下のように規格化因子＝０．４０９と算出される。また、条件１の場合には、クラスタ２の規格化因子も０．４０９と算出される。

$$\text{規格化因子（条件1，クラスタ1）}=MAX\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=MAX\left(\frac{{}_{6}C_{m}\times{}_{6}C_{n}}{{}_{6+6}C_{2+1}}\right)=\frac{{}_{6}C_{2}\times{}_{6}C_{1}}{{}_{12}C_{3}}$$

$$=\frac{15\times6}{220}=0.409 \qquad \cdots(式12)$$

したがって、（式１２）の計算式に条件１を代入した場合における補正項２（１）の値は、補正項２（1）＝１と算出される。同様に、混在クラスタ２の補正項２（１）の値も１と算出される。

したがって、補正項２（1）の値は、下式（式１３）のように１と算出されるので、特に補正は行なわれずに、類似率は０．５と算出される。

$$\text{類似率(式10，条件1)}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}[(\text{補正項1})\times(\text{補正項2})\times(\text{補正項3})\times\delta]$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\left\{1\times\frac{\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}{MAX\left(\frac{{}_{M}C_{x}\times{}_{N}C_{y}}{{}_{M+N}C_{m+n}}\right)^{\varepsilon}}\times1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{0.409}{0.409}+\frac{0.409}{0.409}\right)=0.5 \qquad \cdots(式13)$$

上記の（式１３）により算出される類似率（式１０に条件１を代入した場合）＝0.5の値は、補正を考慮しない（式１）による類似率の演算結果と一致している。そして、技術文献群に含まれる技術文献数量がそれぞれ６個と６個であり、混在クラスタ内に含まれる技術文献数も２個と１個であるので、われわれが常識的に考えた類似の程度とほぼ一致し

ている。したがって、補正項２（１）を挿入した場合であっても許容範囲内の結果を得ることが可能となる。

下式（式１４）に、計算例１０－２（式１０に条件２を代入した場合）の計算結果について説明する。

条件２の場合のクラスタ１に含まれる技術文献の混在確率は、第１の技術文献群（Ａ群）と第２の技術文献群（Ｂ群）の大きさの比率に近いので、類似率を算出する際にはクラスタ１を構成する技術文献の混在比率の影響を重視して、類似率を大きく算出するべきなのは明らかである。

以下の（式１４）に、補正項２（１）の分子を構成する混在確率の計算例を示す。

$$\text{混在確率（条件2，クラスタ1）}=\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{100}\times{}_{55}C_{50}}{{}_{104+55}C_{100+50}}\right)=\frac{{}_{104}C_{100}\times{}_{55}C_{50}}{{}_{159}C_{150}}$$
$$=\frac{4598126\times3478761}{1.42E+14}=0.113 \qquad \text{...(式14)}$$

一方、分母の規格化因子は混在クラスタ１の混在確率の最大値であるので、以下のように規格化因子＝０．２８０と算出される。また、条件２の場合には、クラスタ２の規格化因子も０．２８０と算出される。

$$\text{規格化因子（条件2，クラスタ1）}=MAX\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{98}\times{}_{55}C_{52}}{{}_{104+55}C_{98+52}}\right)$$
$$=\frac{{}_{104}C_{98}\times{}_{55}C_{52}}{{}_{159}C_{150}}=\frac{1.52E+09\times26235}{1.42E+14}$$
$$=0.280 \qquad \text{...(式15)}$$

したがって、条件２におけるクラスタ１の補正項２（１）の値は、補正項２（１）＝0.404 と算出される。また、条件２におけるクラスタ２

の補正項２（１）の値は、「１」と算出されるので、下式（式１６）に示すように、補正項２（１）に基づく類似率は0.351と算出される（図１２参照）。

$$\text{類似率(式10, 条件2)}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ=1}}^{\text{全クラスタ数}}\{(\text{補正項1})\times(\text{補正項2})\times(\text{補正項3})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ=1}}^{\text{全クラスタ数}}\left\{1\times\frac{\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}{MAX\left(\frac{{}_{M}C_{x}\times{}_{N}C_{y}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}\times1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{0.113}{0.280}+\frac{0.448}{0.448}\right)=0.351 \qquad \ldots(式16)$$

上記の（式１６）にて算出した類似率（式１０に条件２を代入した場合）＝0.351の値は、クラスタ１に含まれる技術文献の混在確率に引っ張られ、類似率（式４に条件２を代入した場合）＝0.962から、類似率（式５に条件２を代入した場合）＝0.351に補正された。

下式（式１７）～（式１９）に、計算例１０－３（式１０に条件３を代入した場合）の計算結果について説明する。以下の（式１７）は、補正項２（１）の分子を構成する混在確率の計算例である。

$$\text{混在確率（条件3，クラスタ1）}=\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{20}\times{}_{55}C_{50}}{{}_{104+55}C_{20+50}}\right)=\frac{{}_{104}C_{20}\times{}_{55}C_{50}}{{}_{159}C_{70}}$$

$$=\frac{1.28E+21\times3478761}{1.49E+46}=2.98E-19\fallingdotseq0.000 \qquad \ldots(式17)$$

一方、分母の規格化因子は混在クラスタ１の混在確率の最大値であるので、以下のように規格化因子＝0.133と算出される。また、条件３の場合には、クラスタ２の規格化因子も0.448と算出される。

$$\text{規格化因子(条件3, クラスタ1)} = MAX\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right) = \left(\frac{{}_{104}C_{46}\times{}_{55}C_{24}}{{}_{104+55}C_{46+24}}\right)$$

$$= \frac{{}_{104}C_{46}\times{}_{55}C_{24}}{{}_{159}C_{70}} = \frac{7.96E+29\times 2.49E+15}{1.49E+46}$$

$$= 0.133 \qquad \text{...(式18)}$$

したがって、条件３における補正項２（１）の値は、補正項２（１）＝0.000 と算出される。混在クラスタ２の補正項２（１）の値は、条件１及び条件２の場合と同様に１と算出される。

したがって類似率は、下記の計算によって０．２５と算出される。

$$\text{類似率(式10, 条件3)} = \frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\{(\text{補正項1})\times(\text{補正項2})\times(\text{補正項3})\times\delta\}$$

$$= \frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\left\{1\times\frac{\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}{MAX\left(\frac{{}_{M}C_{x}\times{}_{N}C_{y}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}\times 1\times\delta\right\}$$

$$= \frac{1}{4}\left(\frac{0.000}{0.133}+\frac{0.448}{0.448}\right) = 0.25 \qquad \text{...(式19)}$$

上記の（式１９）にて算出した類似率（式１０に条件３を代入した場合）＝0.25 の値は、クラスタ１に含まれる技術文献の混在確率に引っ張られ、類似率（式４に条件３を代入した場合）＝0.459 から類似率（式１０に条件３を代入した場合）＝０．２５に補正された。

下式（式２０）～（式２４）に、計算例１０－４（式１０に条件４を代入した場合）の計算結果について説明する。

条件４の場合には、条件３の場合とクラスタに含まれる技術文献の量の総和は同じであるが、クラスタ１及びクラスタ２に含まれる技術文献

A群と技術文献B群との割合が極端に不均等である場合である。したがって、混合クラスタに含まれる技術文献数が多いからといって類似率を大きく算出しないことが望ましい。

補正項2（1）の混在クラスタ1の分子を構成する混在確率について算出すると、

$$\text{混在確率（条件4，クラスタ1）}=\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{2}\times{}_{55}C_{50}}{{}_{104+55}C_{2+50}}\right)=\frac{{}_{104}C_{2}\times{}_{55}C_{50}}{{}_{159}C_{52}}$$

$$=\frac{5356\times3478761}{2.98E+42}=6.26E-33\fallingdotseq 0.000 \quad \text{...(式20)}$$

一方、分母の規格化因子は混在クラスタ1の混在確率の最大値であるので、以下のように規格化因子＝0.141と算出される。

$$\text{規格化因子（条件4，クラスタ1）}=MAX\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{34}\times{}_{55}C_{18}}{{}_{104+55}C_{34+18}}\right)$$

$$=\frac{{}_{104}C_{34}\times{}_{55}C_{18}}{{}_{159}C_{52}}=\frac{2.91E+27\times1.44E+14}{2.98E+42}$$

$$=0.141 \quad \text{...(式21)}$$

したがって、条件4における混在クラスタ1の補正項2（1）の値は、補正項2（1）＝0.000と算出される。

一方、混在クラスタ2の補正項2（1）の値は、以下のように補正項2（1）＝0.004と算出される。

$$\text{混在確率（条件4，クラスタ2）}=\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{20}\times{}_{55}C_{1}}{{}_{104+55}C_{20+1}}\right)=\frac{{}_{104}C_{20}\times{}_{55}C_{1}}{{}_{159}C_{21}}$$

$$=\frac{1.28E+21\times55}{8.34E+25}=0.001 \quad \text{...(式22)}$$

混在クラスタ２の分母の規格化因子は、混在クラスタ２の混在確率の最大値であるので、条件４の場合には、以下のように規格化因子＝0.194と算出される。

$$\text{規格化因子（条件４，クラスタ２）}=MAX\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)=\left(\frac{{}_{104}C_{14}\times{}_{55}C_{7}}{{}_{104+55}C_{14+7}}\right)$$

$$=\frac{{}_{104}C_{14}\times{}_{55}C_{7}}{{}_{159}C_{21}}=\frac{7.95E+16\times 2.03E+08}{8.34E+25}$$

$$=0.194 \quad \ldots(式2\text{-}3)$$

したがって類似率は、以下のように０．００１と算出される。

$$\text{類似率(式10, 条件4)}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項1})\times(\text{補正項2})\times(\text{補正項3})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left\{1\times\frac{\left(\frac{{}_{M}C_{m}\times{}_{N}C_{n}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}{MAX\left(\frac{{}_{M}C_{x}\times{}_{N}C_{y}}{{}_{M+N}C_{m+n}}\right)^{\gamma}}\times 1\times\delta\right\}$$

$$=\frac{1}{4}\left(\frac{0.000}{0.141}+\frac{0.001}{0.194}\right)=0.001 \quad \ldots(式24)$$

上記の（式２４）にて算出した類似率（式１０に条件４を代入した場合）＝0.001の値は、クラスタ１及びクラスタ２に含まれる技術文献の混在確率が、技術文献Ａ群と技術文献Ｂ群から無作為に取り出した場合の混在確率の最大値よりもはるかに小さいので、類似率（式４に条件４を代入した場合）＝0.459から類似率（式１０に条件４を代入した場合）＝0.001に補正された。

図１２に、補正項２（１）を採用した場合の類似率算出例（補正項２（１）に条件１～４を代入した場合の計算結果）の図表を示す。

同図に示すように、混合クラスタのうち、技術文献がよく混ざっているクラスタ（混在確率が大きい値を示す条件を備えたクラスタ）では、補正項２（１）の値が大きい値を示していることがわかる。また、技術文献がよく混ざっていないクラスタ（混在確率が小さい値を示す条件を備えたクラスタ）では、補正項２（１）の値がほぼ「０」と、小さい値を示し、算出される類似率の値も小さい値を示している。

図１３に、補正項１（１）及び補正項２（１）の双方を採用した場合の類似率算出例（補正項１（１）及び補正項２（１）に、条件１～４を代入した場合の計算結果）の図表を示す。

条件１の場合に算出された類似率＝0.5は、われわれが常識的に考える類似の程度とほぼ合っているといえる。

条件２の場合には、混合クラスタ１に含まれる技術文献の数量はクラスタ２～４に含まれる技術文献数量と比べると明らかに多いが、算出される類似率の値は（式１）に条件２を代入した場合の類似率＝0.5から、補正項１（１）及び補正項２（１）を用いて条件２を代入した場合の類似率＝0.4に補正された。このように補正項１（１）及び補正項２（１）を用いて類似率を算出することによって、技術文献数の多いクラスタ１についてあまり重み付けをしたくないときに有効となる。

条件３の場合には、条件２の場合と比較してクラスタ内に含まれる技術文献数量の総和は同じだが、混在クラスタ１の技術文献数量のみが特に多い訳ではないので、算出される類似率の値は類似率＝0.019に小さく補正された。このように補正項１（１）及び補正項２（１）を用いて類似率を算出することによって、クラスタ１に含まれる技術文献数量の多さを類似率の算出結果に反映したくない場合に有効となる。

条件４の場合には、条件２の場合と比較してクラスタ内に含まれる技術文献数量の総和は同じだが、混在クラスタ１や混在クラスタ２が特に大きい訳ではなく、技術文献の混ざり具合がさらに極端なとき、類似率の値は類似率＝0.0005に補正されている。このように補正項１（１）及び補正項２（１）を用いて類似率を算出することによって、混在クラス

タ内の技術文献数量が多い場合であっても、技術文献の混ざり具合が不均等である場合には類似率を小さく算出する方向に補正することが可能となる。

すなわち、補正項１（１）及び補正項２（１）を用いて類似率を算出することによって、技術文献数量の多い混在クラスタを重要視して類似率を補正するとともに、技術文献の混ざり具合が不均一な場合には、類似率を小さい値に補正することが可能となる。

また同図に示すように、補正項２（１）の計算式では、補正項の値が技術文献の混ざり具合に敏感に反応する傾向があるので、適宜γの値を調節する必要が生ずる場合もあると考えられる。そして、混在クラスタ内に含まれる技術文献の数量に基づいた補正と、混在クラスタ内に含まれる技術文献の混ざり具合に基づく補正とは、上述のようにそれぞれ密接な関係があるので、αの値とともにγの値を適宜定めることも重要であると考えられる。

なお、図１３はα＝１、γ＝１とした場合の計算例であるが、例えばα＝１のままとしてγ＝0.25に設定して試算してみると、条件１の類似率＝0.5→0.5、条件２の類似率＝0.4→0.769、条件３の類似率＝0.019→0.019、条件４の類似率＝0.0005→0.033と算出することが可能となる。

応用型３：補正項２（２）の算出例

補正項２（２）は、混在クラスタ内における技術文献の混在比に応じて類似率を補正する補正項である。

第１の技術文献群（Ａ群）と、第２の技術文献群（Ｂ群）に含まれる技術文献の数量の比が大きく異なる場合には、各混在クラスタに含まれる技術文献の混在比も当然異なるはずである。また、両群に含まれる技術文献の数量が拮抗しているほど、クラスタに含まれる技術文献の混在比は第１の技術文献群（Ａ群）及び第２の技術文献群（Ｂ群）に含まれ技術文献数の数量の比（構成比）に近くなると考えるのが妥当である。

そこで本発明では、第１の技術文献群（Ａ群）及び第２の技術文献群

（Ｂ群）に含まれる技術文献数の構成比Ｎ／Ｍと、各クラスタ内における技術文献数の混在比ｎ／ｍについて、更に構成比と混在比との比を取ったもののξ乗（但し、０＜ξ）に比例した補正値を、類似率を算出する際の補正項として設けている。

すなわち、第１の技術文献群（Ａ群）及び第２の技術文献群（Ｂ群）に含まれる技術文献数の構成比Ｎ／Ｍと、各クラスタ内における技術文献数の混在比ｎ／ｍが近いほど類似率を高く設定する（１に近づける）ための数式である。

したがって補正項２（２）の値は、第１の技術文献群（Ａ群）及び第２の技術文献群（Ｂ群）に含まれる技術文献数の構成比と、各クラスタ内における技術文献同士の混在比が異なるほど１から小さい値を取る。

$$\text{補正項2(2)}=\frac{\frac{N}{M}\text{又は}\frac{n}{m}\text{の小さい方}}{\frac{N}{M}\text{又は}\frac{n}{m}\text{の大きい方}}=\left\{\frac{MIN\left(\frac{N}{M},\frac{n}{m}\right)}{MAX\left(\frac{N}{M},\frac{n}{m}\right)}\right\}^{\xi}=MIN\left(\frac{N\times m}{M\times n},\frac{M\times n}{N\times m}\right)^{\xi}\quad\text{...(式25)}$$

補正項２（２）を考慮した類似率の算出例を、以下の（式２６）に示す。

$$\text{類似率}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項1})\times(\text{補正項2})\times(\text{補正項3})\times\delta\}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left[(\text{補正項1})\times\left\{\frac{MIN\left(\frac{N}{M},\frac{n}{m}\right)}{MAX\left(\frac{N}{M},\frac{n}{m}\right)}\right\}\times(\text{補正項3})\times\delta\right]\quad\text{...(式26)}$$

上記の（式２５）及び（式２６）に示すように補正項２（２）では、技術文献Ａ群及び技術文献Ｂ群の構成比と、各クラスタ内における技術文献同士の混在比が同じであるほど類似率を高く設定する（1に近づける）ために、分子には「Ｎ／Ｍ又はｎ／ｍの小さい方」を配置し、分母には「Ｎ／Ｍ又はｎ／ｍの大きい方」を配置している。

この場合に、技術文献の混在比が小さい混在クラスタの影響を、類似率の算出結果に大きく反映させたくない場合には、補正項の指数ζを $\zeta > 1$ に設定するとよい。

また、単純にクラスタ内における技術文献の混在比に応じて類似率を増減させる要望がある場合には、$\zeta = 1$ に設定するとよい。

また、混在比が大きい混在クラスタの影響を類似率の算出結果に大きく反映させたくない要求がある場合には、$0 < \zeta < 1$ に設定するとよい。

以下に、類似率の計算に際して補正項２（２）を用いる場合の作用について説明する。

補正項２（２）では、分子にＡ群とＢ群の技術文献数量の構成比又は各クラスタ内における技術文献同士の混在比のいずれか小さい方を配置し、分母にＡ群とＢ群の技術文献数量の構成比又は各クラスタ内における技術文献同士の混在比のいずれか大きい方を配置するようにしたので、Ａ群とＢ群の技術文献数量の構成比と各クラスタ内における技術文献同士の混在比が同じであるほど類似率を高く算出する（1に近づける）ことが可能となる。また、Ａ群とＢ群の技術文献数量の構成比と各クラスタ内における技術文献同士の混在比が異なるほど類似率を小さい値に算出することが可能となる。

また、Ａ群とＢ群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比を算出しているので、類似率の算出範囲を $0 \leq$ 類似率 ≤ 1 を保証することが可能となる。

更に、指数ζを $\zeta > 1$ に設定することによって、Ａ群とＢ群の技

術文献数量の比と、各クラスタ内における技術文献同士の混在比との比が小さい混在クラスタの影響を、類似率の算出結果に大きく反映させないようにすることが可能となる。

また、指数ζを　ζ＝１　に設定することによって、単純にA群とB群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比に応じて類似率を増減させることが可能となる（単純混在比比較）。

また、分子の指数を０＜ζ＜１　に設定することによって、A群とB群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比が大きい場合に類似率の算出結果に対する影響を少なくすることが可能となる。

補正項２（２）のみを考慮して他の補正項の作用を考慮しない場合であって（すなわち補正項１＝１、補正項３＝１とする）、単純混在比比較を行なう場合（すなわちζ＝１）に、技術文献群同士を比較する条件として、（式２６）において条件１～４に設定した場合の類似率の試算結果を以下に示す。なお、算出結果は、図１４に、補正項２（２）を採用した場合の類似率算出例（補正項２（２）に条件１～４を代入した場合の計算結果）の図表として示す。

下式（式２７）に、計算例２６－１（式２６に条件1を代入した場合）の計算結果を示す。

条件１では、第１の技術文献群（Ａ群）の技術文献数量は６個、第２の技術文献群（Ｂ群）の技術文献数量も６個であるので、A群と群B群の技術文献数量の構成比は１対１である。

一方、条件１の場合に各混在クラスタ（クラスタ１及びクラスタ２）に含まれる技術文献数は、第１の技術文献群（Ａ群）の技術文献が２個、第２の技術文献群（Ｂ群）の技術文献が１個であるので、混在比は２対１である。

したがって、クラスタに含まれる技術文献の混在比による類似率の補正の影響は、少なからず存在することが期待される。

$$\text{類似率（式２６，条件１）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left[(\text{補正項１})\times\left\{\frac{MIN\left(\frac{N}{M},\frac{n}{m}\right)}{MAX\left(\frac{N}{M},\frac{n}{m}\right)}\right\}^{\zeta}\times(\text{補正項３})\times\delta\right]$$

$$=\frac{1}{4}\sum_{\text{クラスター}=1}^{4}\left[1\times\left\{\frac{MIN\left(\frac{6}{6},\frac{n}{m}\right)}{MAX\left(\frac{6}{6},\frac{n}{m}\right)}\right\}^{1}\times1\times\delta\right]$$

$$=\frac{1}{4}\left\{\frac{MIN\left(\frac{6}{6},\frac{1}{2}\right)}{MAX\left(\frac{6}{6},\frac{1}{2}\right)}+\frac{MIN\left(\frac{6}{6},\frac{1}{2}\right)}{MAX\left(\frac{6}{6},\frac{1}{2}\right)}\right\}=\frac{1}{4}\left\{\frac{\frac{1}{2}}{\frac{6}{6}}+\frac{\frac{1}{2}}{\frac{6}{6}}\right\}=0.25\ldots(\text{式27})$$

下式（式２８）に、計算例２６＝２（式２６に条件２を代入した場合）の計算結果を示す。

$$\text{類似率（式２６，条件２）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left[(\text{補正項１})\times\left\{\frac{MIN\left(\frac{N}{M},\frac{n}{m}\right)}{MAX\left(\frac{N}{M},\frac{n}{m}\right)}\right\}^{\zeta}\times(\text{補正項３})\times\delta\right]$$

$$=\frac{1}{4}\sum_{\text{クラスター}=1}^{4}\left[1\times\left\{\frac{MIN\left(\frac{55}{104},\frac{n}{m}\right)}{MAX\left(\frac{55}{104},\frac{n}{m}\right)}\right\}^{1}\times1\times\delta\right]$$

$$=\frac{1}{4}\left\{\frac{MIN\left(\frac{55}{104},\frac{50}{100}\right)}{MAX\left(\frac{55}{104},\frac{50}{100}\right)}+\frac{MIN\left(\frac{55}{104},\frac{1}{2}\right)}{MAX\left(\frac{55}{104},\frac{1}{2}\right)}\right\}=\frac{1}{4}\left\{\frac{\frac{50}{100}}{\frac{55}{104}}+\frac{\frac{1}{2}}{\frac{55}{104}}\right\}$$

$$=\frac{1}{4}\left\{\frac{104}{110}+\frac{104}{110}\right\}=\frac{1}{4}\times1.891=0.473\ldots(\text{式28})$$

下式（式２９）に、計算例２６－３（式２６に条件３を代入した場合）の計算結果を示す。

条件３の場合には、条件２の場合とクラスタに含まれる技術文献の量の総和は同じであるが、混在クラスタ１に含まれる技術文献の混在比が、第１の技術文献群（Ａ群）と第２の技術文献群（Ｂ群）の構成比と大きく異なる状況である。したがって類似率を算出する際に、混在クラスタ１に含まれる技術文献の混在比率の影響が条件２の場合ほどは生じないことが望ましい。

$$\text{類似率（式２６，条件３）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスタ}=1}^{\text{全クラスタ数}}\left[(\text{補正項１})\times\left\{\frac{MIN\left(\frac{N}{M},\frac{n}{m}\right)}{MAX\left(\frac{N}{M},\frac{n}{m}\right)}\right\}^{\zeta}\times(\text{補正項３})\times\delta\right]$$

$$=\frac{1}{4}\sum_{\text{クラスタ}=1}^{4}\left[1\times\left\{\frac{MIN\left(\frac{55}{104},\frac{n}{m}\right)}{MAX\left(\frac{55}{104},\frac{n}{m}\right)}\right\}\times1\times\delta\right]$$

$$=\frac{1}{4}\left\{\frac{MIN\left(\frac{55}{104},\frac{50}{20}\right)}{MAX\left(\frac{55}{104},\frac{50}{20}\right)}+\frac{MIN\left(\frac{55}{104},\frac{1}{2}\right)}{MAX\left(\frac{55}{104},\frac{1}{2}\right)}\right\}=\frac{1}{4}\left\{\frac{\frac{55}{104}}{\frac{50}{20}}+\frac{\frac{1}{2}}{\frac{55}{104}}\right\}$$

$$=\frac{1}{4}\left\{\frac{1100}{5200}+\frac{104}{110}\right\}=\frac{1}{4}\times1.156=0.289\quad\text{（式29）}$$

上記の（式２９）にて算出した類似率（式２６に条件３を代入）＝0.289の値は、混在クラスタ１に含まれる技術文献の混在比が、第１の技術文献群（Ａ群）と第２の技術文献群（Ｂ群）の構成比と異なることから、類似率は少なく補正される。

したがって、補正２（２）の演算処理を行なうことによって、混在ク

ラスタに含まれる技術文献の量が多い場合であっても、その技術文献の混在比率に応じて類似率を補正することが可能となる。

下式（式３０）に、計算例２６－４（式２６に条件４を代入した場合）の計算結果を示す。

$$\text{類似率（式２６，条件４）}=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\{(\text{補正項１})\times(\text{補正項２})\times(\text{補正項３})\times\delta\}$$

$$=\frac{1}{\text{全クラスタ数}}\sum_{\text{クラスター}=1}^{\text{全クラスタ数}}\left[(\text{補正項１})\times\left\{\frac{MIN\left(\frac{N}{M},\frac{n}{m}\right)}{MAX\left(\frac{N}{M},\frac{n}{m}\right)}\right\}^{\zeta}\times(\text{補正項３})\times\delta\right]$$

$$=\frac{1}{4}\sum_{\text{クラスター}=1}^{4}\left[1\times\left\{\frac{MIN\left(\frac{55}{104},\frac{n}{m}\right)}{MAX\left(\frac{55}{104},\frac{n}{m}\right)}\right\}^{1}\times1\times\delta\right]$$

$$=\frac{1}{4}\left\{\frac{MIN\left(\frac{55}{104},\frac{50}{2}\right)}{MAX\left(\frac{55}{104},\frac{50}{2}\right)}+\frac{MIN\left(\frac{55}{104},\frac{1}{20}\right)}{MAX\left(\frac{55}{104},\frac{1}{20}\right)}\right\}=\frac{1}{4}\left\{\frac{\frac{55}{104}}{\frac{50}{2}}+\frac{\frac{1}{20}}{\frac{55}{104}}\right\}$$

$$=\frac{1}{4}\left\{\frac{110}{5200}+\frac{104}{1100}\right\}=\frac{1}{4}\times0.116=0.029 \quad \text{...(式３０)}$$

上記の（式３０）にて算出した類似率（式２６に条件４を代入した場合）＝0.029の値は、クラスタ１及びクラスタ２に含まれる技術文献の混在比が極端に不均等であるとともに、混在クラスタ１及び混在クラスタ２の混在比が第１の技術文献群（Ａ群）と第２の技術文献群（Ｂ群）の技術文献数量の構成比と大きく異なるので、類似率が少なく補正される。

図１４に、補正項２（２）を採用した場合の類似率算出例（補正項２（２）に条件１～４を代入した場合の計算結果）の図表を示す。

条件１、条件２における混在クラスタ１及び混在クラスタ２、並びに条件３における混在クラスタ２は、図９に示すように技術文献がよく混ざ

っている状態であるといえる例（混在クラスタにおける技術文献の混在比が、第１の技術文献群と、第２の技術文献群に含まれる技術文献の数量の比に近い場合）である。この場合には、補正項の値を割合に大きく算出し、類似率の値を大きくする効果がある。

逆に、条件３の混在クラスタ１及び条件４の各混在クラスタは、技術文献がよく混ざっていない状態であるといえる（混在クラスタにおける技術文献の混在比が、第１の技術文献群と、第２の技術文献群に含まれる技術文献の数量の比と大きく異なる場合といえる）ので、補正項の値を小さく算出し、類似率を小さい値に算出する効果がある。

したがって（式４）に示したような、補正項１（１）と組み合わせて類似率を算出することによって、技術文献群同士がどの程度技術的に関連があるかを示す類似率の算出制度を向上させることが可能となる。

図１５に、補正項１（１）及び補正項２（２）を採用した場合の類似率算出例（補正項１（１）及び補正項２（２）に条件１～４を代入した場合の計算結果）の図表を示す。

同図に示すように、補正項１（１）及び補正項２（２）を用いた計算式に条件１を代入すると、クラスタ内に含まれる技術文献数量と混在比率に応じた類似率を算出するので、条件１を代入した場合の類似率＝0.25 の値は、（式１）に条件１を代入した場合（補正項なしの場合）の類似率＝0.5 の値より小さいが、かなり期待した値に近く、技術文献群同士の技術の類似性をよく表していると言える。

また、補正項１（１）及び補正項２（２）を用いた計算式に条件２を代入すると、クラスタ内に含まれる技術文献数量と混在比率に応じた類似率を算出するので、類似率は（式１）に条件２を代入した場合（補正なしの場合）の類似率＝0.5 から、補正項１及び補正項２（２）を用いて条件２を代入した場合の類似率＝0.909 に補正され、かなり期待した類似率の値に近く、技術文献同士の類似性をよく表していると言える。

このように補正項１及び補正項２（２）を用いて類似率を算出することによって、技術文献数の多いクラスタ１について重み付けをすること

が可能となる。

また、補正項１（１）及び補正項２（２）を用いた計算式に条件３を代入すると、クラスタ内に含まれる技術文献数量と混在比率に応じた類似率を算出するので、条件２の場合と比較してクラスタ内に含まれる技術文献数量の総和は同じだが、混在クラスタ１の技術文献数量のみが特に多いわけではなく、かつ、クラスタ１内の技術文献の混在比率が第１の技術文献群（Ａ群）と第２の技術文献群（Ｂ群）の技術文献数量の比率とも異なる場合には、クラスタ１の存在を特に重視しないようにすることが可能となる。

ここで算出される類似率は、（式１）に条件３を代入した場合（補正なしの場合）の類似率＝0.5から、補正項１及び補正項２（２）を用いて条件３を代入した場合の類似率＝0.111に補正され、かなり期待した値に近く、技術文献群同士の類似性を表していると言える。

また、補正項１（１）及び補正項２（２）を用いた計算式に条件４を代入すると、クラスタ内に含まれる技術文献数量と混在比率に応じた類似率を算出するので、条件２の場合と比較してクラスタ内に含まれる技術文献数量の総和は同じだが、混在クラスタ１や混在クラスタ２が特に大きいわけではなく、技術文献の混ざり具合がさらに極端な場合で、混在クラスタ内における技術文献の混在比が、Ａ群とＢ群の技術文献数の比と大きく異なるので、類似率に反映する影響が小さくなっている。

ここで算出される類似率は、（式１）に条件４を代入した場合（補正なしの場合）の類似率＝0.5から、補正項１及び補正項２（２）を用いて条件４を代入した場合の類似率＝0.019に補正され、かなり期待した値に近く、技術文献同士の類似性をよく表していると言える。

応用型４：補正項２（３）の算出例

以下に、混在クラスタ内における技術文献の期待値差に基づく補正について説明する。

あるクラスタ内に含まれる第１の技術文献群（Ａ群）の技術文献の数量Ｍと、第２の技術文献群（Ｂ群）の技術文献の数量Ｎとが、Ａ群とＢ

群から無作為に抽出した際の期待値（M／（M＋N））に近いほど、良く混ざっていると考えるのは自然である。（前記（式９）に示した確率比、又は（式２５）に示した混在比と並ぶ第３の混ざり具合の定義である。）

そこで本発明では、第１の技術文献群（A群）と第２の技術文献群（B群）とを混合した技術文献群の中から、第１の技術文献群（A群）の技術文献を取り出す確率（M／（M＋N））に、混在クラスタに含まれる技術文献数（m＋n）を乗算して第１の技術文献群（A群）の技術文献を取り出す期待値を算出し、その期待値と混合クラスタに含まれる第１の技術文献群（A群）の技術文献数mとの差を期待値差（下（式３１）参照）として算出し、この差が小さいほど（０に近いほど）類似率が高くなるように補正する演算を行なう。

以下の（式３１）に期待値差の算出例を示す。

$$\text{期待値差}=\left|(m+n)\left(\frac{M}{M+N}\right)-m\right|$$
$$=\frac{|mM+nM-mM-mN|}{M+N}$$
$$=\frac{|nM-mM|}{M+N}\quad\text{...(式31)}$$

図１６に、上記の（式３１）に条件１～４を代入した場合の期待値差の算出例を示す。

上記の（式３１）による計算結果からもわかるとおり、あるクラスタ内に含まれるA群の技術文献の数量と、B群の技術文献の数量とが、A群とB群から無作為に抽出した際の期待値に近いほど、そのクラスタを重要視して類似率を補正する場合には、図１６に示す期待値差を負の数にして指数部分に置くとよい。

負の値にした期待値差を指数部分に配置することによって、混在クラ

スタに期待値どおりの技術文献が存在する場合には、期待値差＝0となり、指数＝０の場合には、補正項の値を１と算出することが可能となるからである。ところが、期待値のままだと混ざり具合だけでなく所定の混在クラスタの大きさにも依存してしまうため、期待値差をクラスタに含まれる技術文献数で除算するとよい。

このようにして求めた補正項２（３）の実施例を以下に示す。

$$\text{補正項２（３）} = \xi^{-\frac{|nM-mN|}{(M+N)(m+n)}} \quad \text{...(式３２)}$$

但し、

ξ：任意定数であって、ξ＞１とする。

上記（式３２）のように補正項２を算出することによって、例えば、クラスタの大きさが１００で期待値差が１０の時とクラスタの大きさが１０で期待値差が１の時の補正値を同じにすることが可能となる。

なお、ξの値を大きく設定するほど期待値差に対して敏感に反応して類似率を小さく補正することが可能となる。

図１７に、ξ＝１０とした場合において、（式３２）に条件１～４を代入した場合の類似率算出例を示す。

図１８に、補正項１（１）及び補正項２（３）を採用した場合の類似率算出例（補正項１（１）及び補正項２（３）に条件１～４を代入した場合の計算結果）の図表を示す。

同図に示すように、補正項１（１）及び補正項２（３）を用いた計算式に条件１を代入すると、クラスタ内に含まれる技術文献数量と期待値差に応じた類似率を算出する（あるクラスタ内に含まれる第１の技術文献群（Ａ群）の技術文献の数量と、第２の技術文献群（Ｂ群）の技術文献の数量とが、Ａ群とＢ群から無作為に抽出した際の期待値に近い程類似率を大きく算出する補正を行なう）ので、補正項１及び補正項２（３）

を用いて条件1を代入した場合の類似率＝0.340は、(式1)に条件1を代入した場合（補正なしの場合）の類似率＝0.5の値に近く、期待した値に近い値を算出することが可能となっている。

条件2の場合には、混在クラスタ1は、クラスタ2～4と比べると混在クラスタに含まれる技術文献数が大きい上に、期待値差も少ないので混在クラスタ1に含まれる技術文献の構成の影響を重視すべきである。

補正項1（1）及び補正項2（3）を用いた計算式に条件2を代入すると、クラスタ内に含まれる技術文献数量と期待値差に応じた類似率を算出する（あるクラスタ内に含まれる第1の技術文献群（A群）の技術文献の数量と、第2の技術文献群（B群）の技術文献の数量とが、A群とB群から無作為に抽出した際の期待値に近い程類似率を大きく算出する補正を行なう）ので、補正項1及び補正項2（3）を用いて条件2を代入した場合の類似率＝0.935は、（式1）に条件1を代入した場合（補正なしの場合）の類似率＝0.5の値より大きく補正されており、この値は期待した値に近い値となる。

条件3の場合には、前記の条件2の場合と比較してクラスタに含まれる技術文献数量の総和は同じだが、混在クラスタ1だけが特に大きい訳ではないのでクラスタ1を特に重視しないはずである。また、混在クラスタ1に含まれる技術文献は、第1の技術文献群（A群）と第2の技術文献群（B群）から無作為に抽出した際の期待値と大きく異なるので、混在クラスタ1の期待値差の大きさに引っ張られ類似率は小さく算出されるはずである。

補正項1（1）及び補正項2（3）を用いた計算式に条件3を代入すると、クラスタ内に含まれる技術文献数量と期待値差に応じた類似率を算出する（あるクラスタ内に含まれる第1の技術文献群（A群）の技術文献の数量と、第2の技術文献群（B群）の技術文献の数量とが、A群とB群から無作為に抽出した際の期待値に近い程類似率を大きく算出する補正を行なう）ので、補正項1及び補正項2（3）を用いて条件3を代入した場合に、類似率＝0.207と算出される。この類似率の値も期

待した値に近い値である。

条件４の場合には、条件３と比べてクラスタ内に含まれる技術文献数量の総和は同じだが、混在クラスタ１や混在クラスタ２に含まれる技術文献数量が特に大きい訳ではなく、混ざり具合がさらに極端な場合なので、混在クラスタ１の重み付けに引っ張られないことが望ましい。

補正項１（１）及び補正項２（３）を用いた計算式に条件４を代入すると、クラスタ内に含まれる技術文献数量と期待値差に応じた類似率を算出する（あるクラスタ内に含まれる第１の技術文献群（Ａ群）の技術文献の数量と、第２の技術文献群（Ｂ群）の技術文献の数量とが、Ａ群とＢ群から無作為に抽出した際の期待値に近い程類似率を大きく算出する補正を行なう）ので、補正項１及び補正項２（３）を用いて条件４を代入した場合には、類似率＝0.146と算出される。この類似率の値も、期待した値に近い値である。

産業上の利用可能性

本発明によれば、特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との技術的な類似性を判断するための指標を算出する類似率算出装置であって、比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、第１の技術文献群及び第２の技術文献群に含まれる技術文献について前記入力した技術情報を含む技術文献を検索して該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数との比を類似率として算出する類似率算出手段と、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段とを備えたので、その分解した全クラスタ数と混在クラスタ数の比に基づいて、技術文献群に記載されている技術内容の類似性を示す指標を簡便に算出することが可能となる。

また本発明によれば、類似率算出手段に各混在クラスタに含まれる技術文献の量に応じた値を取る第１の補正値と、各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値を取る第２の補正値とを乗算したものを、各混在クラスタについて総和を算出して、全クラスタ数で除算して類似率を算出する機能を設けたので、補正項１の存在により混在クラスタに含まれる技術文献の量に応じて重要度が高いことを意味付ける補正が可能となるとともに、補正項２の存在により混在クラスタに含まれる技術文献の割合が所定の量に近い程、重要なクラスタであるとして、類似率が高い値を示すように重い重み付けをして、類似率の算出結果を、より人の感覚に合うように補正することが可能となる。

従って、補正項１及び補正項２を用いて類似率を算出することによって、技術文献数量の多い混在クラスタを重要視して類似率を補正するとともに、技術文献の混ざり具合が不均一な場合には、類似率を小さい値に補正することが可能となる。

また本発明によれば、類似率算出手段に個々のクラスタ内の技術文献数のα乗（但し、０＜α）に比例した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する機能を設けたので、クラスタ内の技術文献数が多いほど重要なクラスタであるとするような類似率を算出することが可能となる。

また本発明によれば、類似率算出手段に個々のクラスタ内の技術文献数のα乗（但し、０＜α）を、全クラスタ数等の規格化因子で除算して類似率を算出する機能を備えたので、０≦類似率≦１を保証することが可能となる。また、規格化因子として全クラスタ内の技術文献数の平均値を配置したので、全クラスタ内の技術文献数の平均値を基準として技術文献の量の多少を算出することが可能となる。

また本発明によれば類似率算出手段に、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）に比例した補正値を各混在クラスタについて総和を算出

し、全クラスタ数で除算して類似率を算出する機能を設けた。すなわち、類似率算出手段に（Ａ群の中からｍ個、Ｂ群の中からｎ個の技術文献を取り出す組合せの数）／（Ａ群とＢ群とを混ぜ合わせた中からｍ＋ｎ個の技術文献を取り出す組合せ数）を分子に配置した演算を行なう機能を備えたので、混在クラスタに含まれるＡ群及びＢ群の技術文献数の偏り（作為性）に応じて、偏り大の場合は小さい補正値に、偏り小の場合は大きい補正値に類似率を補正することが可能となる。また、規格化因子として、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率の最大値のγ乗（但し、$0 < \gamma$）を配置したので、類似率の算出範囲として$0 \leq$類似率≤ 1を保証することが可能となる。

また本発明によれば類似率算出手段に、第１の技術文献群に含まれる技術文献数Ｍと第２の技術文献群に含まれる技術文献数Ｎとの構成比、Ｎ／Ｍと、クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群の技術文献数ｍと第２の技術文献群の技術文献数ｎの混在比、ｎ／ｍとについて、更に構成比と混在比との比を取ったもののζ乗（但し、$0 < \zeta$）に比例した補正値を各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率を算出する機能を備えたので、Ａ群とＢ群の技術文献数量の構成比と各クラスタ内における技術文献同士の混在比が同じであるほど類似率を高く算出する（１に近づける）ことが可能となる。

また、構成比と混在比との比の指数ζを$\zeta > 1$に設定することによって、Ａ群とＢ群の技術文献数量の比と、各クラスタ内における技術文献同士の混在比との比が小さい混在クラスタの影響を、類似率の算出結果に大きく反映させないようにすることが可能となる。

また、指数ζを$\zeta = 1$に設定することによって、単純にＡ群とＢ群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比に応じて類似率を増減させることが可能となる。

また、分子の指数を$0 < \zeta < 1$に設定することによって、Ａ群とＢ

群の技術文献数量の構成比と、各クラスタ内における技術文献同士の混在比との比が大きい場合に類似率の算出結果に対する影響を少なくすることが可能となる。

また本発明によれば類似率算出手段に、第1の技術文献群と第2の技術文献群とを混合した技術文献群の中から第1の技術文献群の技術文献を取り出す確率に前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第1の技術文献群の技術文献を取り出す期待値を算出し、前記期待値と混合クラスタに含まれる第1の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を任意定数ξ（但し、1＜ξ）の負の指数とした補正値を、各混在クラスタについて総和を算出し、全クラスタ数で除算して類似率と算出するようにしたので、ξの値の設定に応じて期待値差に対する類似率の算出結果を敏感に反応させる補正を行なうことが可能となる。

また本発明によれば類似率算出手段に、第1の技術文献群と第2の技術文献群とを混合した技術文献群の中から第1の技術文献群の技術文献を取り出す確率に前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第1の技術文献群の技術文献を取り出す期待値を算出し、前記期待値と混合クラスタに含まれる第1の技術文献群の技術文献数との差を期待値差として算出し、その期待値差を混在クラスタに含まれる技術文献数で除算したものを、任意定数ξ（但し、1＜ξ）の負の指数とした補正値とし、これを各混在クラスタについて総和を算出し、更に全クラスタ数で除算して類似率と算出するようにしたので、ξの値の設定に応じて期待値差に対する類似率の算出結果を敏感に反応させる補正を行なうことが可能となる。

請求の範囲

1．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数との比を類似率として算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

2．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算

出するとともに、

各混在クラスタに含まれる技術文献の量に応じた値を取る第１の補正値と、各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値を取る第２の補正値とを乗算したものを各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

３．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

個々のクラスタ内の技術文献数のα乗（但し、０＜α）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

４．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出す

る類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

個々のクラスタ内の技術文献数のα乗（但し、$0 < \alpha$）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

５．請求の範囲４に記載の類似率算出装置における前記規格化因子は、全クラスタ内の技術文献数の平均値であることを特徴とする類似率算出装置。

６．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献

群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

７．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

8．請求の範囲7に記載の類似率算出装置における前記規格化因子は、第1の技術文献群の中からm個、第2の技術文献群の中からn個の技術文献を取り出す確率の最大値のγ乗（但し、0＜γ）であることを特徴とする類似率算出装置。

9．特許文献又は技報等の技術文献から構成される第1の技術文献群と第2の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第1の技術文献群及び第2の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第1の技術文献群及び第2の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第1の技術文献群及び第2の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第1の技術文献群に含まれる技術文献数Mと第2の技術文献群に含まれる技術文献数Nとの構成比、N／Mと、

前記クラスタ分解した結果得られた混在クラスタに含まれる第1の技術文献群の技術文献数mと第2の技術文献群の技術文献数nの混在比、n／mとについて、更に構成比と混在比との比を取ったもののζ乗（但し、0＜ζ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

１０．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、

前記期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、

その期待値差を任意定数ξ（但し、１＜ξ）の負の指数とした補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

１１．特許文献又は技報等の技術文献から構成される第１の技術文献群と第２の技術文献群との、技術的な類似性を判断するための指標を算出する類似率算出装置であって、

比較対象となる第１の技術文献群及び第２の技術文献群を入力する

技術文献群入力手段と、

キーワードやＩＰＣなどの技術情報を入力する技術情報入力手段と、

第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解するクラスタ分解手段と、

前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、

前記期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、

その期待値差を混在クラスタに含まれる技術文献数で除算したものを、任意定数ξ（但し、１＜ξ）の負の指数とした補正値とし、これを各混在クラスタについて総和を算出し、更に前記算出した全クラスタ数で除算して類似率を算出する類似率算出手段と、

前記算出した類似率を記録手段、表示手段、又は通信手段に出力する出力手段と、

を備えたことを特徴とする類似率算出装置。

１２．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似

率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数との比を類似率として算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

１３．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第

2の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、各混在クラスタに含まれる技術文献の量に応じた値を取る第１の補正値と、各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値を取る第２の補正値とを乗算したものを各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

１４．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第

２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、個々のクラスタ内の技術文献数のα乗（但し、０＜α）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

１５．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力

する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、個々のクラスタ内の技術文献数のα乗（但し、０＜α）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

１６．請求の範囲１５に記載の類似率算出プログラムにおいて、

前記情報処理手段に、

前記類似率算出手段が、前記規格化因子として、全クラスタ内の技術文献数の平均値を用いる機能を実現させることを特徴とする類似率算出プログラム。

１７．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第

２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する機能と

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

１８．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

１９．請求の範囲１８に記載の類似率算出プログラムにおいて、

前記情報処理手段に、

前記類似率算出手段が、前記規格化因子として、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率の最大値のγ乗（但し、０＜γ）を用いる機能を実現させることを特徴とする類似率算出プログラム。

２０．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数と

を算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、第１の技術文献群に含まれる技術文献数Ｍと第２の技術文献群に含まれる技術文献数Ｎとの構成比、Ｎ／Ｍと、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群の技術文献数ｍと第２の技術文献群の技術文献数ｎの混在比、ｎ／ｍとについて、更に構成比と混在比との比を取ったもののζ乗（但し、０＜ζ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

２１．技術文献群を入力する技術文献群入力手段と、キーワードなどの

技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、

前記期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、

その期待値差を任意定数ξ（但し、１＜ξ）の負の指数とした補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

２２．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段と、前記技術文献群入力手段と技術情報入力手段とクラスタ分解手段と類似率算出手段と出力手段とを制御をすることが可能な情報処理手段とを備えた類似率算出装置の情報処理手段にて動作し、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出プログラムであって、

前記情報処理手段に、

前記技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する機能と、

前記技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する機能と、

前記クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する機能と、

前記類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、

前記期待値と混合クラスタに含まれる第１の技術文献群の技術文献

数との差を期待値差として算出し、

その期待値差を混在クラスタに含まれる技術文献数で除算したものを、任意定数ξ（但し、1＜ξ）の負の指数とした補正値とし、これを各混在クラスタについて総和を算出し、更に前記算出した全クラスタ数で除算して類似率を算出する機能と、

前記出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する機能と、

を実現させることを特徴とする類似率算出プログラム。

２３．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数との比を類似率として算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手

段に出力する工程と、

を含むことを特徴とする類似率算出方法。

２４．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、各混在クラスタに含まれる技術文献の量に応じた値を取る第１の補正値と、各混在クラスタに含まれる第１の技術文献群の技術文献と第２の技術文献群の技術文献との混ざり具合に応じた値を取る第２の補正値とを乗算したものを各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

２５．技術文献群を入力する技術文献群入力手段と、キーワードなどの

技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、個々のクラスタ内の技術文献数のα乗（但し、０＜α）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

２６．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であっ

て、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、個々のクラスタ内の技術文献数のα乗（但し、０＜α）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

２７．請求の範囲２６に記載の類似率算出方法において、

前記類似率算出手段が、規格化因子として、全クラスタ内の技術文献数の平均値を用いる工程を含むことを特徴とする類似率算出方法。

２８．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

２９．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含ま

れる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群及び第２の技術文献群の技術文献数の確率に応じて補正するために、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率のγ乗（但し、０＜γ）を規格化因子で除算した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

３０．請求の範囲２９に記載の類似率算出方法において、

前記類似率算出手段が、規格化因子として、第１の技術文献群の中からｍ個、第２の技術文献群の中からｎ個の技術文献を取り出す確率の最大値のγ乗（但し、０＜γ）を用いる工程を含むことを特徴とする類似率算出方法。

３１．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、第１の技術文献群に含まれる技術文献数Ｍと第２の技術文献群に含まれる技術文献数Ｎとの構成比、Ｎ／Ｍと、前記クラスタ分解した結果得られた混在クラスタに含まれる第１の技術文献群の技術文献数ｍと第２の技術文献群の技術文献数ｎの混在比、ｎ／ｍとについて、更に構成比と混在比との比を取ったもののζ乗（但し、０＜ζ）に比例した補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

３２．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、

期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、

その期待値差を任意定数ξ（但し、１＜ξ）の負の指数とした補正値を各混在クラスタについて総和を算出し、前記算出した全クラスタ数で除算して類似率を算出する工程と、

出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

３３．技術文献群を入力する技術文献群入力手段と、キーワードなどの技術情報を入力する技術情報入力手段と、技術情報群を技術情報毎にクラスタ分解するクラスタ分解手段と、全クラスタ数と混在クラスタ数とを算出して類似率を算出する類似率算出手段と、算出した類似率を出力する出力手段とを備えた類似率算出装置を用いて、技術文献群同士の技術的な類似性を判断するための指標を算出する類似率算出方法であって、

技術文献群入力手段が、比較対象となる第１の技術文献群及び第２の技術文献群を入力する工程と、

技術情報入力手段が、キーワードやＩＰＣなどの技術情報を入力する

工程と、

クラスタ分解手段が、第１の技術文献群及び第２の技術文献群に含まれる技術文献について、前記入力した技術情報を含む技術文献を検索して、該検索した技術文献をそれぞれの技術情報毎にクラスタ分解する工程と、

類似率算出手段が、前記クラスタ分解した結果得られた全クラスタ数と、第１の技術文献群及び第２の技術文献群の双方の技術文献を含む混在クラスタ数を算出するとともに、

第１の技術文献群と第２の技術文献群とを混合した技術文献群の中から、第１の技術文献群の技術文献を取り出す確率に、前記クラスタ分解した混在クラスタに含まれる技術文献数を乗算して第１の技術文献群の技術文献を取り出す期待値を算出し、

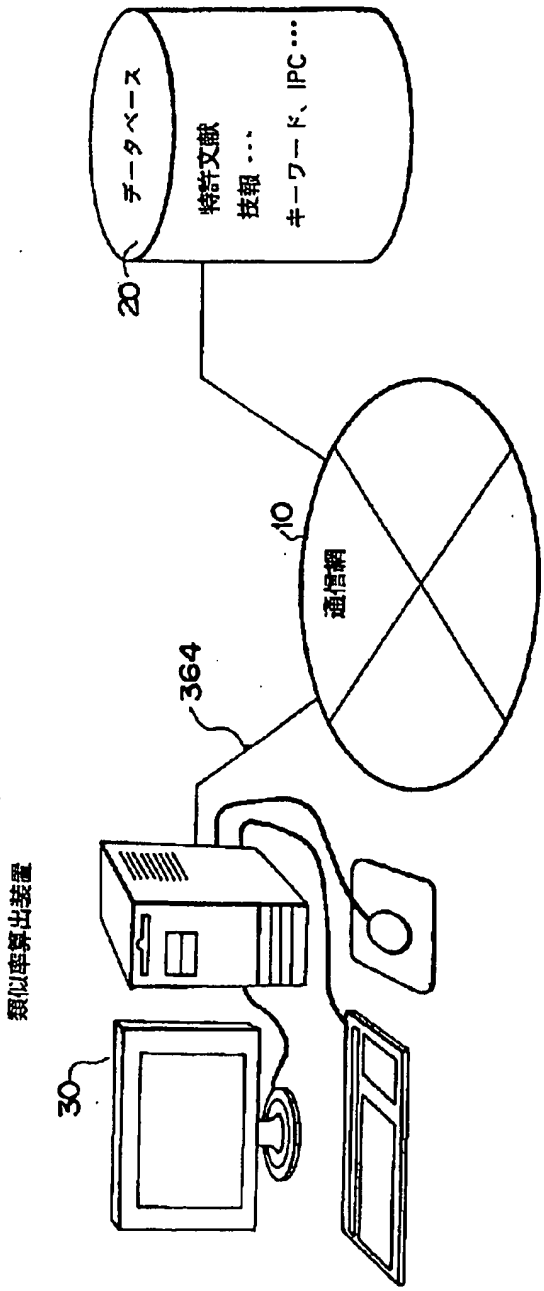
期待値と混合クラスタに含まれる第１の技術文献群の技術文献数との差を期待値差として算出し、

その期待値差を混在クラスタに含まれる技術文献数で除算したものを、任意定数ξ（但し、１＜ξ）の負の指数とした補正値とし、これを各混在クラスタについて総和を算出し、更に前記算出した全クラスタ数で除算して類似率を算出する工程と、

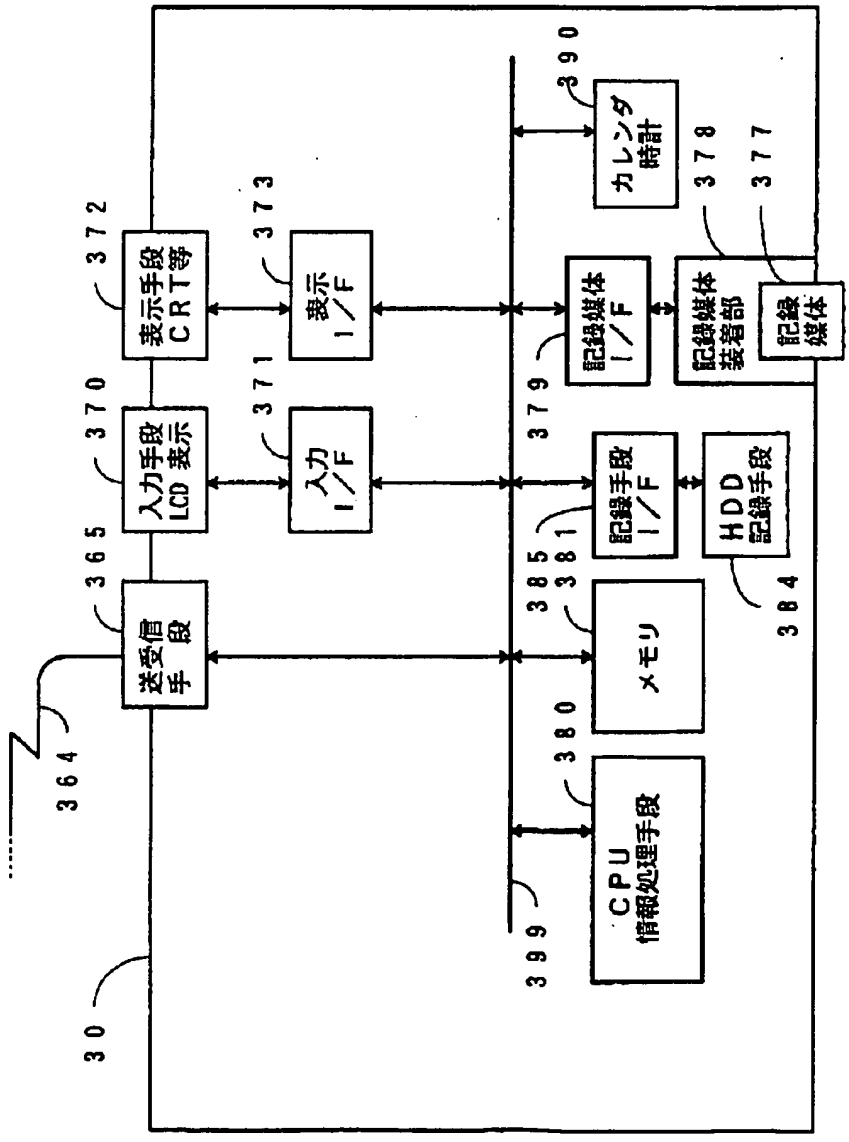
出力手段が、前記算出した類似率を記録手段、表示手段、又は通信手段に出力する工程と、

を含むことを特徴とする類似率算出方法。

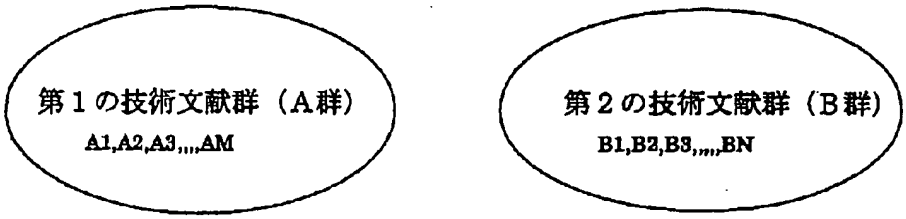
第1図



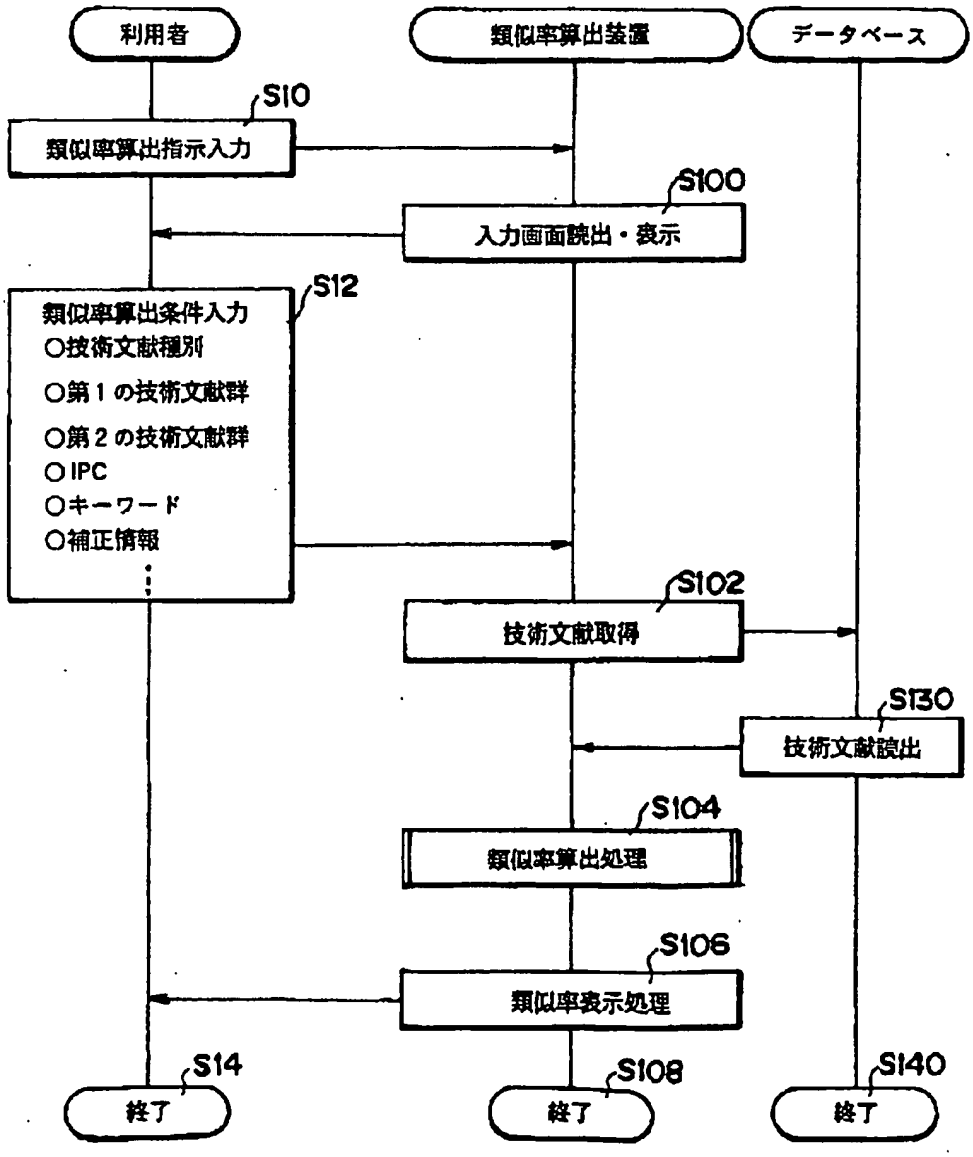
第 2 図


30
364
365
送受信手段
370
入力手段 LCD 表示
372
表示手段 CRT等
371
入力 I/F
373
表示 I/F
379
記録媒体 I/F
381
記録手段 I/F
385
メモリ
380
CPU 情報処理手段
399
390
カレンダ 時計
378
記録媒体 装着部
377
記録 媒体
384
HDD 記録手段

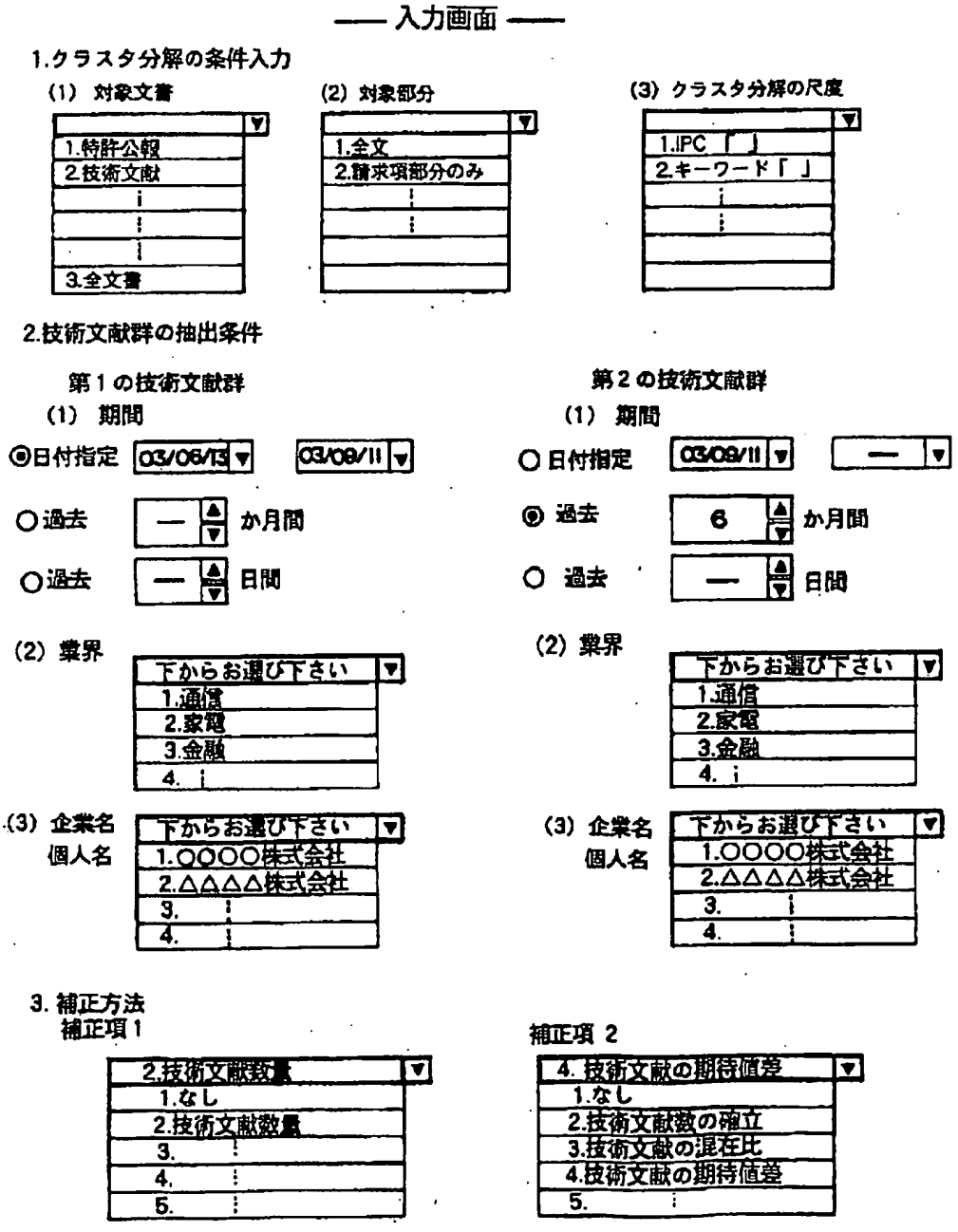
第３図



第 4 図


利用者
類似率算出装置
データベース
S10
類似率算出指示入力
S100
入力画面読出・表示
S12
類似率算出条件入力
○技術文献種別
○第１の技術文献群
○第２の技術文献群
○IPC
○キーワード
○補正情報
S102
技術文献取得
S130
技術文献読出
S104
類似率算出処理
S106
類似率表示処理
S14
終了
S108
終了
S140
終了

第５図


入力画面
1.クラスタ分解の条件入力
(1) 対象文書
1.特許公報
2.技術文献
3.全文書
(2) 対象部分
1.全文
2.請求項部分のみ
(3) クラスタ分解の尺度
1.IPC「 」
2.キーワード「 」
2.技術文献群の抽出条件
第１の技術文献群
(1) 期間
◉日付指定 03/06/13 03/09/11
○過去 — か月間
○過去 — 日間
(2) 業界
下からお選び下さい
1.通信
2.家電
3.金融
4.
(3) 企業名 個人名
下からお選び下さい
1.○○○○株式会社
2.△△△△株式会社
3.
4.
第２の技術文献群
(1) 期間
○日付指定 03/09/11 —
◉過去 6 か月間
○過去 — 日間
(2) 業界
下からお選び下さい
1.通信
2.家電
3.金融
4.
(3) 企業名 個人名
下からお選び下さい
1.○○○○株式会社
2.△△△△株式会社
3.
4.
3. 補正方法
補正項１
2.技術文献数量
1.なし
2.技術文献数量
3.
4.
5.
補正項２
4.技術文献の期待値差
1.なし
2.技術文献数の確立
3.技術文献の混在比
4.技術文献の期待値差
5.

第６図

――類似率表示画面――

1.クラスタ分解の条件：1.特許公報 2.技術文献 7.キーワード「電話」・・・

2.技術文献群の抽出条件

第１の技術文献群

(1) 期間	03／06／13～03／09／11
(2) 業界	1.通信
(3) 企業名	1.○○○○株式会社
(4) その他	

第２の技術文献群

(1) 期間	03／06／13～03／12／13
(2) 業界	2.家電
(3) 企業名	2.△△△△株式会社
(4) その他	

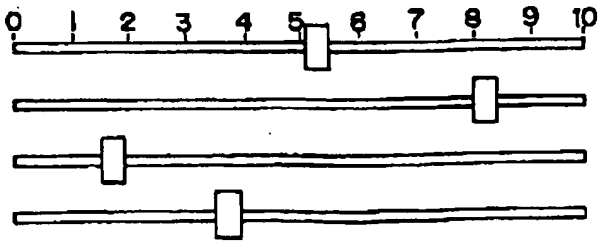
3.補正方法の条件：補正項１＝２ 補正項２＝４ 補正項３＝0.300

4.類似率の算出結果：類似率 0.935

5.分解したクラスタの内容

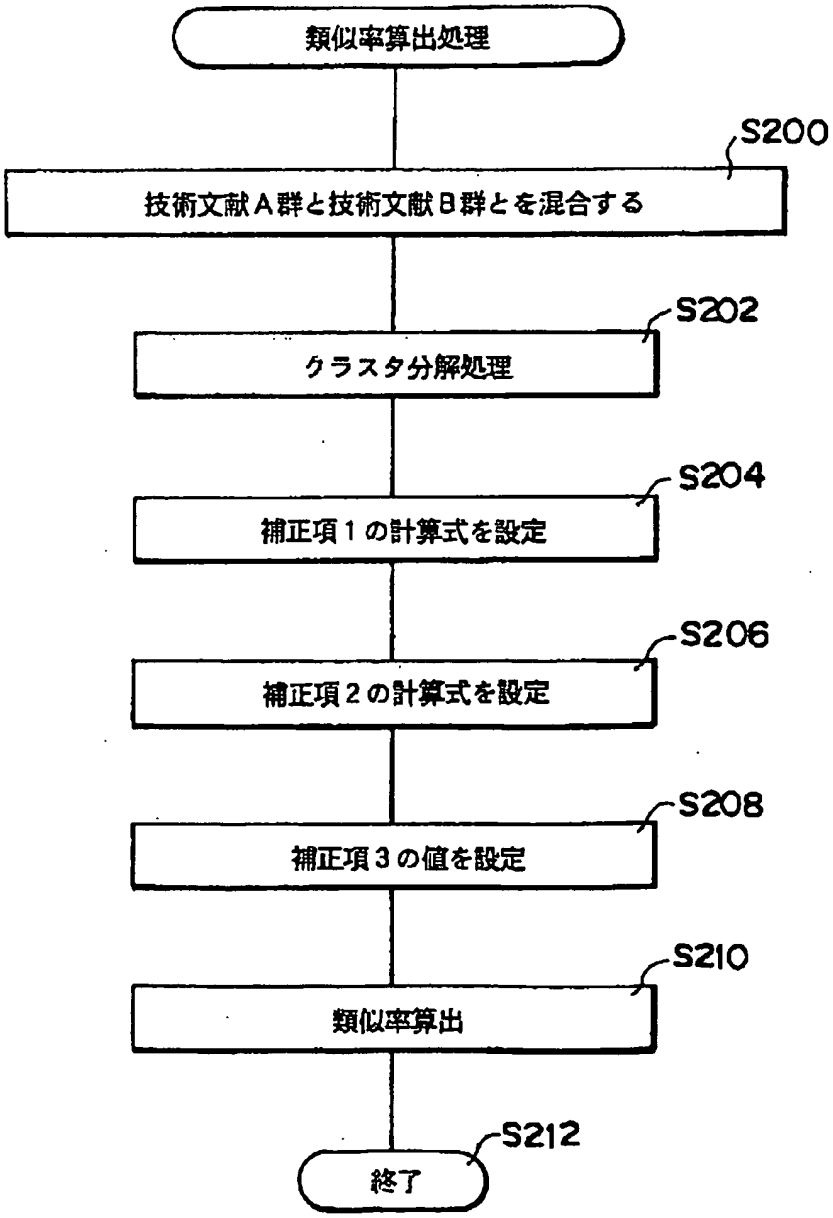
	クラスタ１	クラスタ２	---	---
尺度	「G06F 17／30」	「テキスト処理」	---	---
補正項１	3.774	0.075	---	---
補正項２	0.971	0.971	---	---
補正項３	1.000	1.000	---	---

6.類似率算出条件



第７図

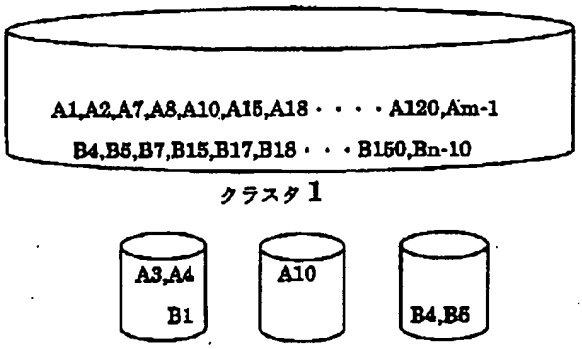
第 8 図


類似率算出処理
S200
技術文献Ａ群と技術文献Ｂ群とを混合する
S202
クラスタ分解処理
S204
補正項１の計算式を設定
S206
補正項２の計算式を設定
S208
補正項３の値を設定
S210
類似率算出
S212
終了

第9図

条件	第1の技術文献群	第2の技術文献群	技術文献合計	クラスタ1			クラスタ2			クラスタ3	クラスタ4	期待する因果値
	技術文献数M	技術文献数N		技術文献数m1	技術文献数n1	技術文献合計	技術文献数M2	技術文献数n2	技術文献合計	技術文献数m3	技術文献数n4	許容範囲±0.050
条件1	6	6	12	2	1	3	2	1	3	2	4	0.300
条件2	104	55	159	100	50	150	2	1	3	2	4	0.900
条件3	104	55	159	20	50	70	2	1	3	82	4	0.200
条件4	104	55	159	2	50	52	20	1	21	82	4	0.050

第１０図


A1,A2,A7,A8,A10,A15,A18・・・・A120,Am-1
B4,B5,B7,B15,B17,B18・・・B150,Bn-10
クラスタ１
A3,A4
B1
A10
B4,B5
クラスタ２ クラスタ３ クラスタ４

第１１図

α＝1、補正項２＝１、補正項３＝１の場合の（式4）による類似率算出例

類似率計算例	(M+N)/4	クラスタ１		クラスタ２		類似率（式４）
		技術文献数 (m1+n1)	補正項１の値	技術文献数 (m2+n2)	補正項１の値	
条件１	12/4	3	1	3	1	0.5
条件２	159/4	150	3.774	8	0.075	0.962
条件３	159/4	70	1.761	3	0.075	0.459
条件４	159/4	52	1.308	21	0.528	0.459

第１２図

γ＝１、補正項１＝１、補正項３＝１の場合の（式１０）による類似率算出例

類似率計算例	クラスタ１			クラスタ２			類似率（式10）
	混在確率	混在確率の最大値	補正項２（１）の値	混在確率	混在確率の最大値	補正項２（１）の値	
条件１	0.409	0.409	1	0.409	0.409	1	0.5
条件２	0.113	0.280	0.404	0.448	0.448	1	0.351
条件３	0.000	0.133	0.000	0.448	0.448	1	0.25
条件４	0.000	0.141	0.000	0.001	0.194	0.004	0.001

第１３図

補正項１（１）及び補正項２（１）を採用した場合の類似率算出例

類似率計算例	クラスタ１			クラスタ２			類似率
	補正項１の値	補正項２(1)の値	補正項１×補正項２(1)	補正項１の値	補正項２(1)の値	補正項１×補正項２(1)	補正項１×補正項２(1)
条件１	1	1	1	1	1	1	0.5
条件２	3.774	0.404	1.525	0.075	1	0.075	0.4
条件３	1.761	0.000	0.000	0.075	1	0.075	0.019
条件４	1.308	0.000	0.000	0.528	0.004	0.002	0.0005

第１４図

ζ＝１、補正項１＝１、補正項３＝１の場合の（式２６）による類似率算出例

類似率計算例	N/M	クラスタ１		クラスタ２		類似率（式２６）
		n1/m1	補正項２（２）の値	n2/m2	補正項２（２）の値	
条件１	1	0.5	0.5	0.5	0.5	0.25
条件２	0.529	0.5	0.945	0.5	0.945	0.473
条件３	0.529	2.5	0.212	0.5	0.945	0.289
条件４	0.529	25	0.021	0.05	0.095	0.029

第１５図

補正項１（１）及び補正項２（２）を採用した場合の類似率算出例

類似率計算例	クラスタ１			クラスタ２			類似率
	補正項１の値	補正項2(2)の値	補正項１×補正項2(2)	補正項１の値	補正項2(2)の値	補正項１×補正項2(2)	補正項１×補正項2(2)
条件１	1	0.5	0.5	1	0.5	0.5	0.25
条件２	3.774	0.945	3.566	0.075	0.945	0.071	0.909
条件３	1.761	0.212	0.373	0.075	0.945	0.071	0.111
条件４	1.308	0.021	0.027	0.528	0.095	0.050	0.019

第１６図

（式３１）に条件１～４を代入した場合の期待値差の算出例

期待値差算出例	クラスタ１			クラスタ２		
	n1×M	m1×N	期待値差	n2×M	m2×N	期待値差
条件１	6	12	0.5	6	12	0.5
条件２	5,200	5,500	1.887	104	110	0.038
条件３	5,200	1,100	25.786	104	110	0.038
条件４	5,200	110	32.013	104	1,100	6.264

第１７図

ξ＝１０、補正項１＝１、補正項３＝１の場合の（式３２）による類似率算出例

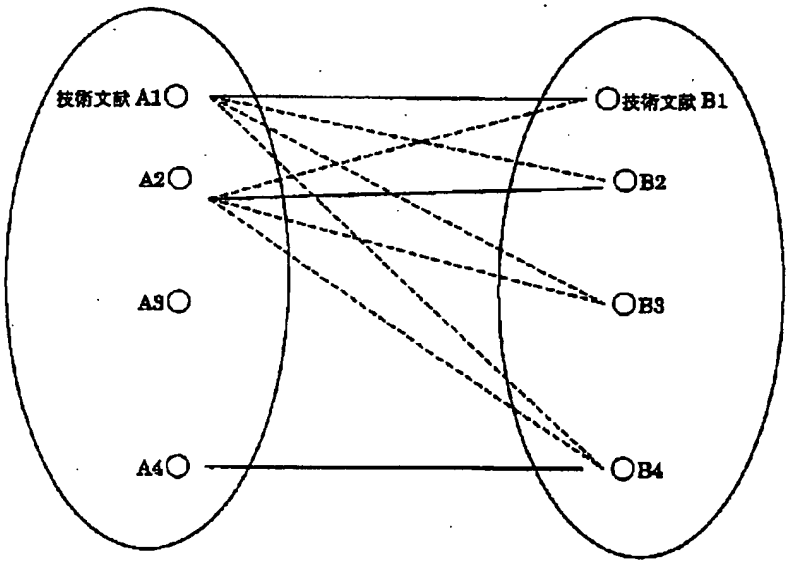
類似率計算例	クラスタ１		クラスタ２		類似率（式３３）
	ξの指数	補正項2(3)の値	ξの指数	補正項2(3)の値	
条件１	0.167	0.681	0.167	0.681	0.340
条件２	0.013	0.971	0.013	0.971	0.485
条件３	0.368	0.429	0.013	0.979	0.350
条件４	0.616	0.242	0.298	0.504	0.187

第１８図

補正項１（１）及び補正項２（３）を採用した場合の類似率算出例

類似率計算例	クラスタ１			クラスタ２			類似率 補正項１×補正項２(3)
	補正項１	補正項２(3)	補正項１×補正項２(3)	補正項１	補正項２(8)	補正項１×補正項２(3)	
条件１	1	0.681	0.681	1	0.681	0.681	0.340
条件２	3.774	0.971	3.665	0.075	0.971	0.073	0.935
条件３	1.761	0.429	0.755	0.075	0.971	0.073	0.207
条件４	1.308	0.242	0.317	0.528	0.504	0.266	0.148

第１９図



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/004451

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Kokai Jitsuyo Shinan Koho 1971-2004
Toroku Jitsuyo Shinan Koho 1994-2004 Jitsuyo Shinan Toroku Koho 1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Kimio ARAI, "Tokkyo Joho Bunseki to Patent Map", The Journal of Information Science and Technology Association, 01 January, 2003 (01.01.03), Vol.53, No.1, pages 16 to 21, full text; all drawings	1-33

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search
28 June, 2004 (28.06.04)

Date of mailing of the international search report
13 July, 2004 (13.07.04)

Name and mailing address of the ISA/
Japanese Patent Office

Facsimile No.

Authorized officer

Telephone No.

Form PCT/ISA/210 (second sheet) (January 2004)

国際調査報告

国際出願番号　PCT／JP2004／004451

A.　発明の属する分野の分類（国際特許分類（IPC））

Int.Cl' G06F17／30

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int.Cl' G06F17／30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国登録実用新案公報	1994－2004年
日本国公開実用新案公報	1971－2004年
日本国実用新案登録公報	1996－2004年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル（JOIS），WPI，INSPEC（DIALOG）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	新井喜美雄，特許情報分析とパテントマップ，情報の科学と技術，2003．1．1，第53巻，第1号，p．16－21，全文，全図	1－33

☐ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日　28．06．2004	国際調査報告の発送日　13．7．2004
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員）　5M　9383 深津　始 電話番号　03－3581－1101　内線　3597

様式PCT／ISA／210（第2ページ）（2004年1月）